DOCUMENT RESUME

ED 213 762 TM 820 207

AUTHOR	Norman, Carol A.
TITLE	Measurement and Testing: An NEA Perspective. NEA Research Memo.
INSTITUTION	National Education Association, Washington, D.C.
PUB DATE	Jul 80
NOTE	80p.
AVAILABLE FROM	National Education Association Professional Library, Box 509, West Haven, CT 06516 ($5.00 NEA members, $10.75 nonmembers).
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	College Entrance Examinations; Educational Policy; *Measurement; *Position Papers; *Professional Associations; Test Coaching; *Testing; Testing Problems; Test Use
IDENTIFIERS	*National Education Association; Scholastic Aptitude Test; Testing Industry; Truth in Testing

ABSTRACT

This Research Memo is for test users, teachers, administrators, counselors, curriculum specialists, school board members, and legislators who use tests as a means for improving education. The purpose of this report is to provide general background information about measurement and testing. The information includes discussions of the meaning attributed to educational measurement, the language of testing, guidelines for test selection, and the uses of test data. A number of problems and issues associated with tests and testing practices are also discussed. The information, together with continued inquiry into tests and testing practices, can help promote informed and responsible use of tests and test data. There are five sections in this report: (1) historical examinations of testing and the developments in psychology and mathematics that helped shape contemporary testing practices and a review of testing practices; (2) coaching for college admission examinations, including data supporting the hypothesis that standardized tests designed to measure aptitude are coachable; (3) results of a National Education Association (NEA) review of commercial involvement in statewide testing programs; (4) review of the NEA position on testing; and (5) truth-in-testing legislation, including state and federal laws. Readers seeking greater detail may find useful the recommended reading list concluding most sections. (Author/GK)

nea helping teachers teach
RESEARCH MEMO
National Education Association
1201 Sixteenth Street, N.W. • Washington, D.C. 20036

Measurement and Testing: An NEA Perspective

MEASUREMENT AND TESTING: AN NEA PERSPECTIVE

Carol A. Norman
Project Director

• *Sale Copies*: NEA members—$5.00 (Stock No. 3080-X-00)
Nonmembers —$10.75 (Stock No. 3080-X-10)

Payment must accompany all orders except for those on your official institutional purchase order forms; no orders can be billed for less than $10. Shipping and handling charges will be added to all billed purchase orders. NEA member discount: 10-49 copies, 10%; 50-99 copies, 15%; 100 or more copies, 20%. Order from the **NEA Distribution Center, The Academic Building, Saw Mill Road, West Haven, CT 06516.**

Contents

FOREWORD

Measurement and testing are central to the educational process and have been with us for many years. The modern era of measurement began in the 1920's and has not changed substantially since that time. This Research Memo serves as a vehicle to communicate NEA's past and present position on standardized testing. Testing resolutions adopted by the NEA Representative Assembly in July 1980 and after the preparation of the body of this report appear in Appendix D. In addition, current testing issues such as coachability of aptitude tests and truth-in-testing legislation are reported.

With the computer assistance of Susan Falsey, Chet McCall and I authored the chapter on the results of the multivariate analysis on the coachability of the SATs. Larry Robinson completed the state survey of commercial involvement in statewide testing programs.

July 1980

Frank W. Kovacs
Director of Research

NEA Research Memo *July 1980*

MEASUREMENT AND TESTING: AN NEA PERSPECTIVE

Introduction

This Research Memo is for test users, that is, for teachers, administrators, counselors, curriculum specialists, school board members, and legislators. Others who may be interested in this report are people who work in elementary and secondary schools, institutions of higher education, and programs for adult learners, and who work with children, teenagers, or adults. Above all, people who use tests as a means for improving education should find this Memo of value.

The purpose of this report is to provide general background information about measurement and testing. The information includes discussions of the meaning attributed to educational measurement, the language of testing, guidelines for test selection, and the uses of test data. A number of problems and issues associated with tests and testing practices are also discussed. This background is not a substitute for excellent textbooks in measurement and testing philosophy, principles of testing, and test evaluation, nor is it intended to guide test development or direct the evaluation of testing programs. Instead, the information, together with continued inquiry into tests and testing practices, can help promote informed and responsible use of tests and test data. In this way the Memo can help people use test results as an aid in improving education.

There are five sections in this report:

I. A historical examination of testing and the developments in psychology and mathematics that helped shape contemporary testing practices. A review of testing practices.

II. The issue of coaching for college admission examinations, including data supporting the hypothesis that standardized tests designed to measure aptitude are coachable.

III. Results of an NEA review of commercial involvement in statewide testing programs.

IV. A review of the NEA position on testing, beginning with the 1972 moratorium and including policy revisions and current policy concerning students, testing, and the instructional process.

V. Truth-in-testing legislation, including state and federal laws. Arguments supporting testing legislation; arguments defending current testing practices; and the NEA position on truth-in-testing legislation, full disclosure, and open testing.

The five sections treat the subjects of measurement and testing broadly. Readers who seek greater detail and who wish to pursue a specific subject on their own may find useful the recommended reading list concluding most sections.

SECTION I: HISTORICAL AND TECHNICAL BACKGROUND

Historical Background

Measurement in education has a long history.[1] As early as 4000 B.C., written examinations were part of the Chinese civil service system and were used to measure how much civil servants had learned. The ancient Spartans had an elaborate series of tests designed to measure mastery of skills of manhood, and in Athens Socrates refined a kind of measure designed to enrich and extend the learning of pupils. During the Middle Ages, the University of Paris introduced the oral examination for master's degree candidates. The practice spread throughout European universities and was extended in 1787 by Frederick William II of Prussia to include secondary students seeking university admission.

In the United States, measurement appeared almost as soon as the first schools were built. A variety of measures were used to assess student achievement, and the oral examination was one of the most popular. It was a quick and easy measure to use, inexpensive, and provided lots of information. It was also controversial. Some students and teachers complained that the exams were unfair and used for punitive reasons. Horace Mann thought so, too; and he successfully argued that written rather than oral exams would provide more objective information.

Tradition and expectations

As measurement is known today, however, it is barely 60 years old. It has during that time become a formal and systematic process complete with theory, a special language, a set of traditions, and what W. James Popham has called a "well established set of expectations."[2] The expectations are that important and sophisticated test development is the task of measurement specialists and that the primary purpose of these tests is to detect diversity among students relative to some measured ability. According to Popham, the expectations are not exactly wrong, for they make sense in light of the history of modern testing and the way tests have been used in the past. Nor are the expectations necessarily right, for they represent a narrow and limiting view of what measurement means, how tests can be used, and the role both measurement and tests can play in education.

Current expectations of educational measurement have a history. If understood, this history can illuminate the present and can give some hint of future change. Therefore, a discussion of current measurement and testing begins properly with the past.

Nineteenth century physiology

During the nineteenth century, two trends began to shape contemporary educational measurement. The first trend in physiology emerged as a group of European scientists began studying human behavior. In 1811, Sir Charles Bell and later Francois Magendie discovered anatomical and functional differences between sensory and motor nerves. This discovery separated nerve physiology into the study of sensation and movement and was followed by a great deal of work with sensation. Some scientists began studying reflex action and what astronomers called the "personal equation," or what is now called reaction time.

The century began with wide acceptance of Immanuel Kant's assertion that psychology could not be experimental. By mid-century some scientists cautiously speculated that the mind might be studied empirically. At this point psychology broke from its affiliation with philosophy and physiology and became known as the new field of experimental psychology.

Individual differences and mental life

Among the interests of early experimental psychologists were the ideas of individual differences and complex mental life and the hypothesis that complex mental life was comprised of a combination of sensory experiences. These ideas had intrigued philosophers, but psychologists tried to measure them. They adapted to their purposes known methods of scientific inquiry and in time established psychological laboratories of which two gained particular prominence: the Wundt laboratory in Germany and the Anthropometric Laboratory established by Sir Francis Galton in England.

Wilhelm Wundt and his associates in Germany studied primarily human sensitivity to sensory stimuli and reaction time, and they searched for uniformity in human behavior. Their legacy to contemporary measurement, however, was methodology. They emphasized rigorous control of experimental conditions and demanded accuracy, precision, order, and the reproducibility of research results, all of which laid the foundation for contemporary measurement standards for objectivity, reliability, and validity.

Sir Francis Galton's initial interest began with the inheritance of genius and led to the measurement of human faculties. Galton is credited with inventing the test as an experimental method; and he and his associates developed a number of tests for studying individuals and, in particular, individual differences.

Thus, the first trend in physiology began to shape into scientific form the study of human behavior. Among the contributions of these early scientists were a theoretical approach to the study of human behavior, the concepts of human attributes whereby individual similarities and differences could be studied, an array of instruments for recording human behavior, and a rigorous methodology.

While physiologists were concerned with sensation and movement, a second trend that influenced contemporary measurement emerged in mathematics. There was during the early nineteenth century considerable interest in observational and instrumental error. The interest was pronounced in astronomy where perfect observations were essential for calibration of the clock. During the 1820s, Friedrich Wilhelm Bessel investigated observational error and discovered personal errors of observation among astronomers observing the same event and by an individual observing across time. Bessel presented the observer differences as a mathematical equation, and efforts were made by astronomers to determine these "personal equations" and to correct for them.

Normal curve and correlation

In 1733, Abraham de Moivre formulated a mathematical theory of error called the theory of probability; Pierre Simon Laplace and Karl Friedrich Gauss demonstrated its usefulness as a mathematical tool early during the eighteenth century; and in 1846, Lambert Quetelet applied it successfully to the measurement of human attributes. This application led to what is known as the normal curve. It is a mathematical model representing the expected distribution of some variable when an infinite number of observations is made. Inspired by this success, Sir Francis Galton, who was studying individual differences and experimenting with tests of mental ability, began to explore ways of applying mathematics to human measurement. One of the concepts he worked out was statistical correlation. His contemporary, Karl Pearson, derived the mathematical formulation for the concept.

By the turn of the century, a number of psychological concepts and mathematical formulations had been developed in Europe for the study of human behavior. When James K. Cattell returned to the United States after doctorate study in Germany, he brought with him many of these concepts and tools. Cattell established psychological laboratories at the University of Pennsylvania and Columbia where he stressed experimental rigor. He was also interested in Galton's work with individual differences and testing. He began developing what he called mental tests and eventually administered them to entering freshmen at Pennsylvania and Columbia. The tests were of various sensory attributes such as reaction time and visual acuity.

For decades psychologists had approached mental measurement and complex mental life through the study of simple sensory experiences. This was the theoretical orientation of Cattell, and it was the approach challenged by Alfred Binet. Binet believed complex mental ability could be measured directly and that the concept itself could be reduced to a number of specific abilities. Binet and Theophile Simon set out to measure complex ability directly and for that purpose developed a scale which they revised and refined over several years. Binet then constructed a formal numerical base for translating test performance into mathematical language. Thus, he succeeded in developing a measure of the characteristic he called intelligence in such a way that the characteristic could be tested.

Army Alpha

Binet's work had profound significance for educational measurement, but it alone was not responsible for the "testing boom" that occurred later. The work of Binet and others in France had been carefully followed by psychologists in the United States. When World War I broke out, the military needed to assemble a huge army quickly and wanted some objective way to classify and place new recruits. Psychologists reasoned that if they could adapt Binet's tests of individuals to groups, they could provide that objective procedure. Their efforts resulted in group tests of mental ability called Army Alpha and Army Beta. The project demonstrated the feasibility of testing many people quickly and simultaneously and was regarded at that time as enormously successful.

After the war, the idea of group tests took hold. A number of tests were constructed, and most were fashioned after the Army Alpha. Most, like Army Alpha, focused on some mental attribute, were of the paper-and-pencil type, and were designed to differentiate among people. Regarded as technically complex and scientific instruments, the tests were also subject to special protection. The belief was that if tests were available to untrained and irresponsible people, they would be misused and spoiled.

Testing boom

Few people questioned publicly the assumptions underlying the new tests, although the tests themselves and more often their use were debated publicly.[3] Nevertheless, many tests were constructed and eventually used in the schools. The schools provided a logical setting for the new tests. A tradition of testing already existed, educators showed interest in the new measures; and the tests provided an attractive alternative to the methods then in use. Mass production made many tests available at reasonable cost. Electronic scoring and computer processing made possible the analysis of data with unprecedented accuracy and speed. General public support of the testing movement and, eventually, publicly issued testing mandates all helped make testing a common educational practice.

Technical Background: The Meaning of Measurement

In one sense, *theory* is a symbolic representation of experience.[4] It is a way of making sense of experience. With theory the logic of experience is reconstructed so

that experience can be contemplated, interpreted, criticized, and unified. Theories are modified as new and unanticipated data are found. They are discarded when they are no longer consistent with the data. New theories take the place of discarded theories and help continue the discovery of generalizations about experience. Thus, theory guides inquiry; it helps explain the present; and it also has the quality of tentativeness.

Mental traits as abstract attributes

The meaning of measurement is provided by theory, and in education prevalent theories concern cognition, learning, and instructional practice. For this reason, the measurement of mental traits has many implications for teaching and learning.

A mental trait is an hypothesized attribute of people. G.C. Helmstadter describes a trait as an abstract attribute.[5] It is not concrete in the sense that it can be known through the senses. One such trait is ability. It cannot be directly seen, heard, touched, or smelled. The trait cannot be known directly at the concrete level of experience. It is an abstraction and is postulated theoretically as an attribute of people. If ability is an attribute important to the learning process, then its measurement would provide information useful to instruction. The problem is how to measure something that has no concrete form.

In theory the trait is presumed to exist and to manifest itself in certain forms of observable human behavior. The task becomes one of identifying those behaviors assumed to reflect the trait, defining the trait so that it can be measured, and constructing an instrument powerful enough to assess the behavior of interest. A test, then, can be thought of as a way of obtaining examples of human behavior. The examples are given a numerical value assumed to resemble the measure of the real underlying trait. Hence, a measurement has been made.

Obviously, measurement is not an end in itself. Its scientific value can be best appreciated as an instrument leading to action. With the assumption that the theory guiding measurement is appropriate, the meaning of measurement is derived from the ends it is intended to serve, the role it is called upon to play, and the functions it performs in inquiry.

Measurement and description

One function of measurement is to refine descriptions and make them more precise. Using numbers to represent traits and their properties allows minute distinctions to be made between observed similarities and differences. With precision, classification systems emerge and ambiguity dissolves to the extent that knowledge permits. Precision does not mean that disagreements are impossible. Disagreements continue to exist; but they are sharpened, at times refocused, sometimes dissolved.

Measurement and decision making

A second function of measurement is its practical utility. The meaning of measurement is often associated with the way it is used and the ends it helps achieve. Since tests in education are functionally viewed as providing information for decision making, the actual functions of testing are referred to in terms of the kinds of decisions they serve.

Lee J. Cronbach uses a three-category system to talk about the practical function of tests in terms of their decision-making function.[6] The categories are research; evaluation; and selection, classification, and placement decisions.

In research an investigator may be interested in the hypothesis that no relationship exists between coaching as a form of preparation to take a certain test and student performance on that test. The investigator may use that test in an experiment designed to test the hypothesis. Use of the test would help the researcher decide whether to accept or reject the hypothesis.

Evaluation is a second kind of situation for which tests are used. Here various measures of the results of a specific training or educational program are obtained and provided to various audiences. The purpose of the measures is to provide evidence in the process of judging the worth or merit of the program. If, for example, a curriculum committee is evaluating a reading program, test scores may be used as one kind of information to help committee members judge the program's worth.

A third kind of situation involves selection, classification, and placement. In selection decisions, some individuals are chosen by preference from among others. Selection implies rejection as some people are chosen and others are not. Admitting students to law school is a selection decision for which tests are used.

Classification decisions involve a systematic arrangement of individuals for prospective treatment. On the basis of a music performance test, for example, students might be classified as beginning, intermediate, or advanced students. Based on this classification, students receive different instructional treatment.

Placement is a special kind of classification. Like classification, placement implies different treatment for different people. Unlike classification where differential treatment is temporary, placement involves relatively long-term differential treatment. One such placement decision occurs when individuals are classified on the basis of some measured trait and then placed in differing instructional programs. The programs often continue for the duration of a student's elementary or secondary school experience, and there is little student movement from one program to another.

Another way of classifying educational decisions focuses on the decision maker and on the power implied by the act of making a decision. From this perspective, decisions can be classified as institutional or individual. Institutional decisions serve institutional needs. They are made in a centralized setting, generally involve policies and guidelines, and may involve categories of students with identical or similar characteristics. A college admissions committee, for example, may make admission decisions about several thousand applicants and may in the process use test data to help identify preferred students. This would be considered an institutional decision.

Individual decisions serve individual rather than institutional needs or preferences. Individual decisions typically take place in decentralized settings. The individual may use information from tests and from other people. The decision is a matter of individual choice, and the power implied in decision making rests with the individual.

Predictive aspect of decision making

There are other ways of classifying the functional use of testing, but the two systems described above are common. The systems focus on the decision making function. Regardless of the kind of decision made, Lee J. Cronbach argues that all decisions involve prediction. According to Cronbach, a test might provide interesting information about individual differences, but this fact might not be worth knowing if one could not predict that these same individuals will differ in some future moment with respect to the same or some other measure. For example, committee members gather information to evaluate a reading program. They examine the evidence and determine that the program is effective and should be continued. Implicit in the decision is the prediction that program effectiveness will continue given the same or similar circumstances. The university requires a test score as part of the application for admission procedure. The test score may be used as the basis for predicting whether students will successfully complete their first year in college. The teacher

administers a music performance test. The prediction here is that certain skill levels can be best developed by certain instructional treatment.

Prediction as a measurement function

The predictive aspect of decision making, of testing, and of the way traits are conceptualized is important. Decisions are made with the expectation that certain desired outcomes are likely to occur. Tests are administered with the expectation that resulting information will improve the predictive dimension of decision making. From the practical point of view, prediction is a function of measurement, and the predictive power of a test is an important characteristic.

Technical Background: The Language of Testing

The term prediction is one of many words associated with the language of testing. It is associated with one of the more technical aspects of testing to be discussed on the following pages. First, it might be useful to consider the way tests are classifed and then to discuss technical characteristics and standards for evaluating tests.

Classifying tests by kind of measure

Tests are classified generally along one of two dimensions: the way traits are measured and the kind of trait measured. Consider first the way traits are measured. Common descriptors are objective and subjective tests, standardized tests, norm-referenced and criterion-referenced tests. The meaning of many of these tests is obvious, but some discussion might be appropriate. Group tests are administered to many individuals simultaneously, but they can be given to single individuals if necessary. Individual tests usually require the manipulation of apparatus and careful questioning or observation and must be administered to one student at a time. All tests can be regarded as requiring some kind of performance, but a performance test usually refers to a task requiring no verbal response.

The terms objective and subjective refer to scoring procedures. A test is subjective if scoring involves judgment on the part of the person doing the scoring. An oral examination is considered quite subjective as are many essay tests. A test is said to be objective if the scoring can be replicated exactly and the same score can be derived regardless of who scores the test and when. A multiple-choice test where each item has only one agreed-upon correct answer can be quite objective. Even with objective tests, however, there is such a thing as scoring error.

Standardized tests are those in which procedures, test materials, and scoring are fixed precisely so that duplication is possible at varying times and places. Standardization was one of the early procedures developed for testing. Without it, results from different experimental laboratories could not be compared. Later standardized classroom tests enabled people to compare test results across classrooms, schools, and regions.

Norm-referenced and criterion-referenced tests involve different standards for interpretation. A criterion-referenced test is one designed to describe a person's score or level of performance in terms of the kind of knowledge, skill, or task he or she can accomplish. A norm-referenced test is one designed to describe performance in terms of a person's relative standing among others who have taken the same test.

Classifying tests by trait

Tests can also be classified according to the trait measured. The broadest classification distinguishes between maximum performance and typical performance. Maximum performance tests are designed to measure a person's best performance. Measurement assumptions are that the test actually brings out best performance

and that the subject is motivated to earn the best possible score. The following tests are traditionally considered measures of maximum performance:

Tests of maximum performance

- Intelligence is sometimes defined as what the test measures, and intelligence tests generally measure verbal, nonverbal, memory, and problem-solving skills. In this sense intelligence tests are often considered measures of general or scholastic aptitude.
- Aptitude tests theoretically measure mental operations that improve little with practice. They provide the basis for predicting levels of future performance.
- Ability tests theoretically measure functions that reflect both innate ability and the influence of general environmental enrichment.
- Achievement tests are designed to measure skills, knowledge, and degree of accomplishment or competence acquired through some educational or training experience.

Tests of typical performance

Tests of typical performance are designed to measure how a person reacts, feels, or behaves. Presumably the test has the power to solicit a sample of characteristic behavior, and the subject is encouraged to demonstrate such behavior. Personality tests, interest inventories, and projective techniques are all examples of typical performance tests.

Tests have a number of other characteristics. They are comprised of one or a number of structured items or exercises. The item provides a performance stimulus and structures the response. For purposes of analysis, an item is the basic scorable unit of a test.

A test is comprised of a number of items which the individual attempts to answer. Each answer is classified according to some numerical scale, usually a two-category scale of right (=1) or wrong (=0). These numbers are called item scores. Item scores are then summed for a given test to yield a raw score which is the total number of items right.

Describing test performance

When a test is given to a number of people, their raw scores can be tallied and described in various ways. Commonly the scores are described by a frequency distribution (the number of people obtaining each score), the cumulative frequency distribution (the number of people who obtained a given score or lower), or such summary statistics as the average (arithmetic mean) and some measure of variability (the range or standard deviation).

The distribution of test scores is influenced primarily by two item characteristics called item difficulty and homogeneity. The difficulty of an item refers to the proportion of students in a given population or sample who got the item right. Thus, an item with a *p* value of 90 means that 90 percent of the students who answered that item gave the right answer. A *p* value of 90 suggests a fairly easy item. A *p* value of 25 suggests a fairly difficult item. Item difficulty affects the mean score.

The homogeneity of an item refers to its correlation with other items in the test. An item may have correlations ranging from a perfect negative correlation with another item (-1) through no correlation at all (0) to a perfect positive correlation (+1). The average correlation between all possible pairs of test items is called a measure of homogeneity. Both item intercorrelations and item difficulty affect variability.

The distribution of test scores is affected by errors of measurement, and measurement errors are detected through their effect on item difficulty and the item intercorrelations. There are errors in any measurement, and test developers try to reduce them. But errors exist, and they keep a test from rendering perfectly valid results.

Validity

Error of measurement is an aspect of a technical characteristic called validity. Validity means truthfulness. To be valid, a test must measure accurately what it is supposed to measure. Validity is the single most important characteristic of a test.

There are three types of validity. Content validity is the degree to which test item content explicitly matches the purpose for which the test is to be used. Content validity usually involves a logical analysis of what the test contains. A group of experts may examine test content and agree that the content samples the domain of knowledge or skill in question. Teachers may examine test content to determine the match between test content, instructional content, and instructional objectives.

A second type of validity is construct validity. Construct validity involves gathering evidence to demonstrate that the theoretical trait measured by the test can be verified experimentally. If two tests measure the same trait, then student performance on the two tests should be more highly correlated than with performance on a third test designed to measure a different trait. A study of construct validity would demonstrate whether this were true or not.

The third type of validity is criterion-related validity (either concurrent or predictive). Criterion-related validity is the degree of accuracy with which a test score can predict a person's performance on some criterion such as performance on another test. If both the predictor and the criterion measures are gathered at the same time, the study is said to be concurrent. If the measures are gathered at different times, then the study is said to be predictive. In both concurrent and predictive validation studies, performance on one measure is used to predict performance on another measure.

Reliability

A second technical characteristic of tests is reliability. Reliability means the consistency of a measure over time. Reliability is the extent to which a test is free of errors when a person is measured more than once by the same instrument or when one administration of a test yields small errors from one test taker to another. A test with high reliability is one that will yield much the same score results for individuals and a group of people under different conditions or situations.

People interested in learning about the technical aspects of educational measurement will find many available materials ranging from introductory texts to technical analyses of testing issues and problems. The readings recommended at the conclusion of this section were selected to represent that range of available materials.

Standards for Test Evaluation

Thousands of tests on the market are available for school use. Their selection should be based on a thorough understanding of the educational purpose of the test and test quality. Among the many considerations involved in test selection, standards of practicality, technical characteristics, and cost-benefit are of primary importance.

Practical considerations

Considerations of *practicality* involve planning and what some people call common sense. The following questions are among those useful for determining the practicality of a test:

- Who will be tested?
- How will individuals be tested, when, and where?
- Who will administer the test?
- Are test procedures feasible? (Consider available space and time, qualifications required of test administrators, ease of administration and scoring, and characteristics of the people being tested.)
- What decision-making purpose will the test data serve?
- Who will make the decision?
- What are the information needs of the decision maker?
- Will the test provide data relevant to the decision needs?
- Will the test provide data important to the decision purpose?
- Will use of the test provide timely data for the decision purpose?
- Who will be affected by the decision?
- In what way will individuals or groups be affected?

Technical considerations

The following questions are useful for determining the technical characteristics of a test:

Standardization

- How is the test administered?
- How is the sample selected for the norming population?
- Who was included in the norming population?
- What are the limitations of the derived scores?

Objectivity

- What method of scoring is used?
- Is the scoring system free of error?

Reliability

- How is reliability determined?
- What is the estimate of reliability for the test?
- What is the standard error of measurement for the test?

Validity

- Does the test have validity for the situation in which it is being used?
- Does the test measure the information and/or performance on an important set of tasks?
- Does the test measure current performance when compared to a standard or criterion measure?
- Does the test measure future performance when compared to a standard or criterion measure?
- Does the test measure a trait or set of characteristics?
- Can an experimental condition be created to test the hypothesis?

Cost-benefit consideration

Questions useful for determining *cost-benefit* include the following considerations:

- What is the material cost of measuring each student?
- What are the service costs involved in testing—e.g., computer scoring, interpretive manuals?
- What are the personnel costs involved in test administration?
- Can the test be locally scored?
- How much useful information will the test provide?
- Can other tests provide the same or better information and at what cost?
- Do budget allocations for testing allow purchase of this test?
- Is the test reusable?
- Can the test serve multiple decision purposes?
- Will the test provide quality information?

FOOTNOTES

[1]Information about the history of educational measurement and the contributions of psychology and mathematics was drawn from Boring, Edwin G. *A History of Experimental Psychology* (2nd ed.) New York: Appleton-Century-Crofts, 1957; Chaucey, H., and Dobbin, J.E. "Testing Has a History." In *Readings in Educational and Psychological Measurement*. C.I. Chase and H.G. Ludlow (Eds.). Boston: Houghton Mifflin Co., 1966; Ebel, Robert L. *Essentials of Educational Measurement*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972; Helmstadter, G.C. *Principles of Psychological Measurement*. New York: Appleton-Century-Crofts, 1964; and Linden, Kathryn W., and Linden, James D. *Modern Mental Measurement: A Historical Perspective*. Boston: Houghton Mifflin Co., 1968.

[2]Popham, W. James. "Educational Measurement for the Improvement of Instruction." *Phi Delta Kappan* 61:531; April 1980.

[3]Cronbach, Lee J. "Five Decades of Public Controversy Over Mental Testing." *American Psychologist* 30:1-14; January 1975.

[4]Kaplan, Abraham. *The Conduct of Inquiry*. San Francisco: Chandler Publishing Co., 1964, p. 294.

[5]Helmstadter, G.C. *op. cit.*, p. 17.

[6]Cronbach, Lee J. *Essentials of Psychological Testing* (3rd ed.). New York: Harper & Row, 1970, pp. 23-25.

[7]*Ibid.* p. 22.

RECOMMENDED READINGS

Carmines, E.G. and Zeller, R.A. *Reliability and Validity Assessment*. Beverly Hills, California: Sage Publications, 1979.

Cronbach, L.J. *Essentials of Psychological Testing* (3rd ed.). New York: Harper & Row, 1970.

Ebel, R.L. *Essentials of Educational Measurement*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

Helmstadter, G.C. *Principles of Psychological Measurement*. New York: Appleton-Century-Crofts, 1964.

Lyman, H.B. *Test Scores and What They Mean* (3rd ed.). Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1978.

Thorndike, R.B., ed. *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1970.

SECTION II: THE ISSUE OF SAT COACHABILITY

Introduction

The Scholastic Aptitude Test (SAT) sponsored by the College Examination Board (CEB)[1] and produced by Educational Testing Service (ETS) is one of the most prominent aptitude tests currently available. With a history spanning more than 50 years, the SAT is an educational measure familiar to millions of people who in the past have taken the test or have used the test results for research, evaluation, and/or selection, placement, or classification decisions. The test has been administered to junior and senior high school students and has been used by colleges as part of the college admission process. Thus, the test is not only familiar but also important to students who seek college entry and to college officials who grant entry privileges.

The SAT has also gained prominence because of the claims made for its power. As a test of aptitude, the SAT is alleged to be a measure of the capacity to learn and is believed, therefore, to be impervious to special preparation or coaching. As a test of scholastic aptitude, the SAT has also been promoted as an indicator of success in college and in particular of college grades earned by students during the first year.

Coaching issue

ETS and CEB have repeatedly claimed that the SAT is not coachable and that it functions as a measure of college success. They have promoted these claims in their publications and have cited various studies in support of their position. In turn, the claims have been challenged, most recently by Warner Slack and Douglas Porter in an article entitled "The Scholastic Aptitude Test: A Critical Appraisal" published by the *Harvard Educational Review*. In this article the authors examine ETS and CEB claims made for the SAT in light of available data.

Slack and Porter begin their appraisal with the 1968 CEB position paper entitled "Effects of Coaching on Scholastic Aptitude Test Scores." In this paper, the stated position is that intensive drill or special tutoring will not significantly increase SAT performance.[2] In support of this position, seven studies are referred to in which the SAT was administered to students before and after some form of coaching.

Slack and Porter examine the seven studies noted by CEB. Contrary to the CEB contention, the authors argue that the seven studies do provide evidence that coaching for the SAT can lead to statistically significant score changes.[3] They speculate that the best available coaching methods were not used in the studies cited by CEB and proceed to examine other studies in which well-planned coaching programs for the SAT resulted in score gains exceeding any gain that could be expected from practice and growth alone.[4]

The authors also examine studies published prior to the 1968 CEB publication on coaching.[5] Of the 29 studies found, 23 were either conducted or sponsored by ETS or CEB. The remaining 6 studies were the result of independent effort. Of the 23 ETS or CEB studies, 15 had been cited by ETS prior to the 1968 publication on coaching. The weighted mean score gains for these 15 studies was 16. The weighted mean score gains of studies not reported by ETS was 41. The weighted mean score gains of all studies found, including those cited by ETS, was 29. The authors

conclude that research evidence does not support the claim that the SAT is uncoachable. Furthermore, the authors argue that this evidence was available prior to the 1968 CEB publication on coaching.[6]

Slack and Porter continue their appraisal by examining data concerning the predictive power of the SAT. The focus of most of these studies is on correlations between SAT test scores and first-year college grades. The power of the SAT is expressed as a validity coefficient and indicates the degree to which SAT scores predict college grades.

Beginning with a predictive validity study conducted with the first SATs in 1926, the authors report that the median validity coefficient for student SAT scores was 0.34, for the high school record for these same students it was 0.52, and for the SAT scores and high school records combined it was 0.55. In 1926, SAT scores increased by 0.03 the predictive power provided by high school records alone.[7] The authors appear skeptical that an increase of 0.03 represents much improvement over the predictive power of high school records.[8]

High school record is best predictor

Slack and Porter examined studies cited in an article by W.B. Schrader entitled "The Predictive Validity of College Board Admissions Tests" in a publication by CEB and validity coefficient tables published by CEB for 1964 through 1974. Using two different methods of interpretation, also employed by ETS, the authors conclude that relatively recent studies show that the SAT adds little to the predictive power of the high school record when considered alone.[9] Thus, as early as 1926 and certainly during the ten-year interval of 1964 through 1974, SAT data provided no evidence that the SAT was a successful indicator of college success.

Based on their critical review of SAT data, Slack and Porter conclude that the SAT is a standardized test of achievement, not a test of aptitude.[10] As a test of achievement, SAT content is far removed from most high school and college curricula.[11] Furthermore, the authors believe performance can be improved with coaching.[12]

The findings reported by Slack and Porter are supported by NEA analyses of SAT data obtained through the Federal Trade Commission. Described in Parts A and B of this section, NEA examined SAT data for evidence of the influence of coaching on average SAT scores, and the degree to which coaching improves SAT scores.

FOOTNOTES

[1]The College Examination Board (CEB) was formerly the College Entrance Examination board (CEEB).

[2]Slack, W.V. and Porter, Douglas. "The Scholastic Aptitude Test: A Critical Appraisal." *Harvard Educational Review* 50:156. May 1980.

[3]*Ibid.*, p. 158.

[4]*Ibid.*, p. 164.

[5]*Ibid.*, p. 161.

[6]*Ibid.*, p. 161.

[7]*Ibid.*, p. 165.

[8]*Ibid.*, p. 165.

[9]*Ibid.*, p. 166-67.

[10]*Ibid.*, p. 169.

[11]*Ibid.*, p. 172.

[12]*Ibid.*, p. 155.

PART A: NEA ANALYSIS OF SAT DATA CONCERNING DIFFERENCES BETWEEN COACHED AND NONCOACHED STUDENTS

FTC finds SAT is coachable

World-wide, special schools exist to prepare or coach individuals to take the SAT. The existence of these schools and their effectiveness have been questioned and in 1976 were investigated by the Federal Trade Commission. The purpose of the FTC study was to provide support for the contention that coaching does improve an individual's test score on the SAT and the Law School Admission Test (LSAT). Based on an analysis of SAT and LSAT data, the FTC reported in 1978 that coaching was dramatically effective for the SAT and that the LSAT was susceptible to coaching. A caveat issued by the Commission, however, explained that some of the conclusions in the study were not supported by evidence obtained from the investigation.

Much of the concern giving rise to this warning centered around the fact that the characteristics of the coached and noncoached groups included in the study were not reasonably similar. For example, the coached group ranked higher in high school, had parents with higher average incomes, included fewer black students, included students with better grades in English and math, and included a greater proportion of students from nonpublic schools (44.7 percent of the coached students attended nonpublic schools whereas only 24.6 percent of noncoached students did so). Thus, score differences could not be attributed to coaching alone.

NEA investigates group differences

In 1979 a revised FTC statistical analysis of SAT data was released. According to this analysis, coaching was found to be effective for students who did not score well on standardized tests, but the initial concern with group differences had not been addressed. Because of its interest in testing and coaching, NEA Research requested and eventually received a copy of the original data tape used to generate the 1978 and 1979 FTC analyses. NEA Research then proceeded to duplicate the statistical analyses reported in the 1979 FTC report. The question addressed in the NEA independent analysis of the FTC data tape was whether differences in SAT scores could be found after taking into consideration some of the possible differences among the coached and noncoached groups of persons (differences such as income level and high school grades).

NEA designed two approaches for grouping persons included in the FTC data base. The first approach involved an effort to match persons in coached and noncoached groups with respect to characteristics which might have an impact on SAT scores and which might influence SAT score differences. Sex, age, family income, and high school standing were among the characteristics considered. Within the available FTC data base, this approach failed to yield a sufficient set of "matched pairs" for meaningful analysis. A second procedure was consequently used.

The second procedure involved two statistical techniques for grouping persons included in the FTC data base and for assessing differences between SAT scores of coached and noncoached groups. The first technique was a discriminatory analysis to determine whether certain characteristics available for some students could be successfully used to classify individuals into coached or noncoached categories. Given that certain characteristics could be identified, the second technique used was an analysis of covariance to determine whether, on the average, a significant difference (better than 90 percent confidence) existed between coached and noncoached SAT test scores after eliminating the linear effects of the characteristics included from the discriminatory analysis.

Given the second approach, four subgroups of high school students, also used in the 1979 FTC report, were identified. With this analysis, coached and noncoached

students could be grouped by grade level, year, and whether they had taken the SAT once or twice. Table 1 presents a descriptive summary of the four subgroups identified for analytical purposes.

Table 1: First- and Second-Time SAT Examinees Grouped by Grade Level, Year, and Coached Status

High School Group	Coached	Noncoached
SAT Examinees: First Time		
Juniors, 1975	50	451
Juniors, 1976	100	487
SAT Examinees: Second Time		
Seniors, 1975	51	305
Seniors, 1976	111	324

In each of the above four groups there was a maximum of 24 characteristics in the data base. For example, parent's income level, high school rank, sex, and latest math grade were reported for all four groups. On the other hand, Preliminary Scholastic Aptitude Test (PSAT) scores were reported for only the first two groups.

The first of the statistical techniques employed considered the available characteristics associated with coached and noncoached students within each of the four groups. The basic question which the technique asked was: Which of the characteristics, if any, can be used to differentiate a student as being in the coached or noncoached category? Putting it another way: Are there certain characteristics which serve to classify a student as coached or noncoached?

Research questions

If such characteristics could be identified, then the next question was: Taking into consideration characteristics which serve to differentiate between the coached and noncoached students, is there a difference in the average SAT scores?

Both of these questions are addressed in the material which follows, for each of the four groups separately. Then a summary is presented, based upon findings for each group.

Juniors, 1975, First Time Takers

Fourteen variables were included in the analysis with all 14 allowing for 74 percent of the students being correctly classified as either coached or noncoached. When criteria for inclusion in the analysis were stipulated, 8 of the 14 were considered and 74 percent correctly identified. The significant variables for discrimination, in the order selected were:

1. Parent's income level
2. Latest math grade
3. PSAT verbal score
4. SAT verbal score
5. High school class rank

6. High school (public or private)
7. PSAT math score
8. SAT math score

There were two analyses of covariance performed on the verbal SAT and the math SAT.

Verbal Score—First Exam:

Significant difference found

Two analyses were performed, the first including the PSAT verbal score and the second excluding the PSAT verbal score (PSAT math score and SAT math score first exam were also excluded for this analysis). There is a significant difference between average verbal scores of coached and noncoached at the 0.02 level. With the PSAT verbal score removed, there is still a significant difference between the mean values.

Math Score—First Exam:

Significant difference found

The same two analyses were performed for the SAT math scores, with and without the PSAT math (SAT verbal and PSAT verbal were not included). The results are identical with the mean differences being significant at better than the 0.01 level in both cases.

Juniors, 1976, First Time Takers

Sixteen variables were included in the analysis with all 16 allowing for 72 percent correctly being classified as coached or noncoached. When criteria for inclusion in the analysis were stipulated, 13 of the 16 were considered and 77 percent correctly classified. Of these 13 only the first 8 were selected for the covariance analyses which follow. The first 8 variables, in the order selected, were:

1. Parent's income level
2. SAT verbal score
3. PSAT verbal score
4. Two PSATs taken before the first SAT
5. Latest math grade
6. PSAT math score
7. SAT math score
8. Sex

There were two analyses of covariance performed on the verbal SAT and math SAT.

Verbal Score—First Exam:

Significant difference found

Two analyses were performed, the first including the PSAT verbal and the second excluding it. There is a significant difference at the 0.01 level between the coached and noncoached average verbal scores both with and without the PSAT verbal being included.

Math Score—First Exam:

The same two analyses were performed for the SAT math scores, both with and without the PSAT math as a factor. With the PSAT math included, there is a significant difference between coached and noncoached at the 0.01 level. With the PSAT math eliminated, there is no significant difference between the coached and noncoached average math scores at the 0.10 level, although average SAT scores for coached are higher than noncoached.

Some difference found

Seniors, 1975, Second Time Takers

Fifteen variables were included in the analysis with all 15 allowing for 72 percent of the students being correctly classified as coached or noncoached. Seven variables accounted for the most variability and were selected in the following order:

1. SAT verbal score—first exam
2. SAT verbal score—second exam
3. Latest math grade
4. Years of English in high school
5. SAT math score—first exam
6. Parent's income level
7. SAT math score—second exam

There were two analyses of covariance performed on the verbal SAT and math SAT where the second exam scores were to be compared.

Verbal Score—Second Exam:

Two analyses were performed, the first including the SAT verbal score (first exam) and the second excluding this variable. In both instances there is a highly significant difference (better than 0.01) between coached and noncoached average grades after eliminating the linear effect of the appropriate variables above.

Significant difference found

Math Score—Second Exam:

The same two analyses were performed for the SAT math score (second exam), both with and without the SAT math score for the first exam. With the first SAT exam score included, there is a significant difference at better than 0.01. With the first SAT math score not included, the difference is significant at the 0.10 level.

Significant difference found

Seniors, 1976, Second Time Takers

When all 15 variables are included in the analysis, 73 percent of the students are correctly assigned to the coached and noncoached groups. With the first 6 variables, the same percentages are properly classified. The 6 significant variables are:

1. SAT verbal score—first exam.
2. SAT verbal score—second exam
3. Latest English grade
4. SAT math score—first exam
5. SAT math score—second exam
6. Parent's income level

There were two analyses of covariance performed on the verbal SAT and math SAT where the second exam scores were compared.

Verbal Score—Second Exam:

Some difference found

Two analyses were performed, with and without the SAT verbal score (first exam) as a factor. With the SAT verbal score on the first attempt as a factor, there is a significant difference between the two groups (better than 0.01). With the SAT verbal score (first exam) excluded, there is no significant difference at the 0.10 level, although average SAT scores are higher than the noncoached.

Math Score—Second Exam:

Significant difference found

The two analyses in this case yield the same conclusions, namely, with or without the math score (first exam) as a factor there is a significant difference between the two groups at better than the 0.01 level.

Conclusions

Based upon the analyses just described, the following conclusions can be drawn:

1. In examining the means and standard deviations for the variables in the analyses, it was evident that the averages for the coached group tended to be higher while the standard deviations tended to be lower than the noncoached group, suggesting not only higher levels of the variables for the coached group but also more homogeneity among the students within the coached group.

Coached and noncoached students differ

2. In all instances, significant discriminations were noted between the coached and noncoached groups. From 73 percent to 77 percent of the individuals in the four groups were correctly assigned to either the coached or noncoached group.

3. Of the 16 analyses performed on the differences between the SAT scores after eliminating the effects of characteristics identified (see Table 2), significant differences exist in 14 of the 16 cases. This finding strongly suggests that differences still exist in the average SAT scores between the coached and noncoached. In all instances, the differential SAT averages were still higher for the coached than the noncoached group. Parent's income level was the one characteristic (external to the SAT and PSAT scores) which appeared as a significantly discriminating variable in all 16 analyses.

Table 2. Summary Results: Levels of Significance Between Coached and Noncoached Students, Classified by Grade Level and Year for Analysis of Covariance[1]

STUDENT GROUP	VERBAL SAT		MATH SAT	
	PSAT Included	No PSAT	PSAT Included	No PSAT
1st Time Takers:				
Juniors 1975	0.02	0.10	0.01	0.01
Juniors 1976	0.01	0.01	0.01	N.S.[2]
	1st SAT Included	No 1st SAT	1st SAT Included	No 1st SAT
Coached between 1st and 2nd SAT:				
Seniors 1975	0.01	0.01	0.01	0.10
Seniors 1976	0.01	N.S.[2]	0.01	0.01

[1]For differences between average SAT scores to be considered as significant, the requirement was 0.10 or less. Results are reported as 0.10, 0.02, and 0.01.
[2]Not significant, but average coached SAT scores were higher than noncoached.

PART B: SHOULD HIGH SCHOOLS COACH STUDENTS TO IMPROVE SAT SCORES?

How much difference does coaching make?

The Federal Trade Commission's reports (1978, 1979) on the coachability of the SAT did not adequately answer the question, "Should high schools coach students to improve SAT scores?" The statistical evidence and related conclusions concerning the coached groups described in the previous part of this section suggest that the answer should be affirmative. There is a more important question, however, that remains unanswered. It is: "To what extent does coaching improve SAT scores?" NEA Research attempted to answer this question by analyzing the FTC data for those students who had taken the Preliminary Scholastic Aptitude Test (PSAT) once and the Scholastic Aptitude Test (SAT) twice.

Method and Procedure

Sample

The FTC data base was compiled for the interval of October 1974 through December 1976. The base included 2,286 SAT coached and 1,777 FTC *matched* but uncoached individuals. The selection of students for the NEA analysis was based on two criteria. First, only students who had taken the PSAT once and the SAT twice would be included in the sample. Second, coached students included in the sample would be drawn from the coaching school identified by previous FTC analyses to have produced the *best* results.

Given these criteria, a sample of 1,324 students was drawn. Of these students, 625 took the examinations in 1975, and 699 students took the examinations in 1976. Students were then placed into one of three groups. Group placement was based on whether students had been coached and, if coached, when coaching had occurred. Because students for the years 1975 and 1976 were considered separately, a total of six subgroups, three subgroups for each year, were ultimately identified. The distribution of students for the six subgroups was as follows:

1975 — 65 students coached between PSAT and first SAT
— 105 students coached between first and second SAT
— 405 noncoached students

1976 — 118 students coached between PSAT and first SAT
— 173 students coached between first and second SAT
— 408 noncoached students.

Statistical Treatment

Three research questions

The purpose of the analysis was to determine individual growth of each group. The individual was used as her or his own control, and no comparisons among groups were made. The measure of growth was the average gain between the PSAT and the second SAT. For each group, three questions were addressed:

1. What was the average gain between the PSAT and the second SAT?
2. Was the average gain statistically significant?
3. Was the average gain practically significant?

The design used for this study and its limitations are discussed extensively by Campbell and Stanley.[1] The design is commonly designated as the one-group pretest-posttest design or as a before-after study with a single group. The statistical

procedure used to measure the average gain was a "t" test for correlated data. Means, standard deviations, and tests of significance were performed for each of the six subgroups on the SAT verbal, SAT math, and total SAT scores.

Tests of significance were performed on the verbal, mathematical, and total gain score for all six subgroups. Three groups for 1975 and three groups for 1976 were analyzed.

Results

Differences between PSAT and second SAT scores were found for each of the three subgroups for 1975 and 1976. Differences were found for all groups on the SAT verbal, SAT math, and total SAT scores. Differences on all of the measures were significant at the 0.01 level. Tables 2 through 7, which appear at the conclusion of this section, present the average gain scores for SAT total, SAT verbal, and SAT math scores for each of the three groups for 1975 and for 1976. For discussion purposes, select data from Tables 2 through 7 have been assembled for Table 1 presented below.

Significant difference found

Table 1. Select Average SAT Point Gain Scores for Coached and Noncoached PSAT and Two-Time SAT Takers for 1975 and 1976

Year		Total Average Point Gain	Verbal Average Point Gain	Math Average Point Gain	Average Family Income
1976	Coached between PSAT and 1st SAT	143	73	70	$29,000
1976	Coached between 1st SAT and 2nd SAT	135	65	70	26,000
1976	Noncoached	60	29	31	21,000
1975	Coached between PSAT and 1st SAT	114	55	59	30,000
1975	Coached between 1st SAT and 2nd SAT	104	47	57	26,000
1975	Noncoached	44	17	27	20,000

According to the data in Table 1, the 1976 group coached between the PSAT and the first SAT achieved the greatest total average gain (143 points). The 1976 group coached between the first and second SAT had an average gain of 135 points, while the 1976 noncoached group had an average gain of 60 points.

The 1975 average gains for all three groups were somewhat smaller. The group coached between PSAT and the first SAT showed an average gain of 114 points. The group coached between the first and second administration of the SAT showed an average gain of 104 points. The noncoached group average gain was 44 points.

The average gain scores for all three groups for both 1975 and 1976 suggest that taking the test three times produces positive results. This was true for the verbal and

math scores as well as for the total scores. It is interesting to note that a relationship appears to exist between coaching and average family income.

Discussion

Coaching improves SAT scores

The results of the analysis of the two coached groups further suggest an affirmative answer to the question, "Does coaching improve SAT scores?" This result is not really too profound if one believes that instruction works. Common sense would suggest that if one is taught four hours a week for ten weeks, there should be on the average some positive results. The extent of the increases appears to be both statistically and practically significant.

ETS on the other hand continues to hold to a position that discourages the use of coaching for the SAT. In a message released by ETS in the early months of 1980 entitled *Accountability, Fairness, and Quality in Testing*, the following information was reported about coaching:

> ETS has stated that our research shows a relatively small average gain in scores as an expected effect from short-term instruction. But this will have little, if any, effect on admissions decisions at most institutions. For instance, ETS research (which has been openly reported) shows the effect of coaching on the Scholastic Aptitude Test (SAT) has typically been a gain of less than 15 points for the verbal section and of less than 20 points for the mathematics section on the SAT scale, which ranges from 200 to 800 points. This corresponds to just two additional items correct per section. When the coaching was restricted to an explanation of the test's item formats and practice with them, research studies report smaller effects. Coaching that incorporates formal instruction in the subject matter with special test preparation has been shown to yield somewhat larger effects, perhaps three additional items correct rather than two. Finally, there is no research evidence to prove the claim that coaching can particularly benefit students from minority groups or students with low initial scores. Students who score low the first time they take a test often show greater gains upon retesting simply as a statistical artifact, without regard to coaching.[2]

In its 1979 report the FTC found that "coaching was effective at the two schools contributing on the average approximately 25 points to students' scores on both the verbal and math SAT exams."[3] The report goes on to say that "the students who attended the effective school (School A) tended to be underachievers on standardized exams, i.e., they scored lower on standardized exams than would have been predicted given their personal and demographic characteristics—grades in school and class ranks."[4]

In brief, the 1979 report of the FTC concerning coachability concluded that "underachievers on standardized exams" could on the average increase their score 25 points on the verbal and 25 points on the math SAT exam with coaching from one of the two schools studied.

The question of cost

NEA believes that if 25 or 50 points (on a 200-800 scale) can make the difference in an admission decision to an undergraduate or graduate school of the student's choice, then parents, students, and educators must decide whether the outcome is worth the $200, $300, or even a $500 coaching expenditure. This type expenditure can be afforded by some parents, but not all families have the wherewithal to invest in a coaching school.

Furthermore, NEA believes all students should have an opportunity to receive coaching free of charge. Unfortunately, the groups (i.e., lower socioeconomic, minorities, and women) who have historically lost the most in achieving opportunities for continued higher education are the same groups which score lower on the SAT. The consequences of the lower test scores in many instances have led to the perpetuation of discrimination against these groups.

NEA supports coaching

The conflicting signals coming from ETS and CEB, along with the FTC's unwillingness first to publish reports on the coachability question in a timely manner and then later to turn over the necessary data to permit an independent analysis, have left the estimated 1.5 million 1979-80 student SAT-takers with no real answer on the coachability question. Consequently, the status quo has been maintained except for the student SAT-takers in New York state. The state's truth-in-testing law requires disclosure of results of the individual student's test, answer sheet, and related right or wrong response.

Coaching and College Admission

The New York truth-in-testing legislation helped escalate the issue of the SAT's coachability. In response to the law and the coachability question, the College Board developed a series of press releases and information memoranda about the coachability of the SAT. These memos, released by the College Board in the fall of 1979, further confounded the issue. For example, according to Stephen H. Ivens, director of program services at the College Board, the "SAT measures reasoning abilities which are developed over time both in and out of school."[5] Ivens goes on to say, "If verbal and mathematical reasoning can be learned, we assume that they can be taught, directly or indirectly."[6] In contrast to this observation, the coaching schools make subtle and blatant claims about their ability to develop these mental operations. In addition, they speak to test-taking skills that include efficient use of time limits, methods used to answer questions, and techniques for successful guessing.

CEB response

Unfortunately, these skills and mental abilities are held to be important by college admissions personnel. Although most admissions officers claim that the SAT and LSAT are not used, by themselves, to admit or reject an applicant, very few deny that the scores play an important part in the admissions process.

In an effort to learn how important SAT scores were for college admission, NEA queried a representative sample of eight universities about SAT score admission requirements (See Table 9 at the end of this section). One university indicated that there was no minimum for the verbal and math score; however, if the total was not 1,000 (out of a maximum of 1,600), there was no need to apply. Two other institutions had ranges of acceptable scores (i.e., 500-800, 450-600) for both the verbal and math as well as the total score. The remaining five schools all had minimum scores for admission on all parts and the total.

When this information about admission requirements and the results of the FTC data analyses on the coached and noncoached students who took the PSAT and SAT twice are assessed, recommendations about coaching become obvious. The average total gain scores for the 1976 groups were Group A, coached between the PSAT and first SAT, 143; Group B, coached between the first and second SAT, 135; and Group C, noncoached, 60. A summary of the average gain on the three groups' verbal, math, and total SAT scores appears in Table 8 at the end of this section.

When the results of the average gain effects are compared to the SAT admission requirements of the eight universities shown in Table 9, the following information is revealed:

- Groups A, B, and C on the average *would not* be eligible for admission to any of the eight universities, based on the average converted PSAT scores.
- Groups A and B on the average *would be* eligible for admission to *four* of the *eight* universities, based on the second SAT average scores for the verbal, math, and total.
- Group C on the average *would be* eligible for admission to *one* of the *eight* universities, based on second SAT average scores for the total and subtotals.

The differences between becoming eligible for admission to one versus four of the eight representative universities leaves little doubt that on the average coaching does have a positive effect not only on improving the scores but on increasing the possibility of being favorably considered for admission.

Conclusions

A statistical analysis was performed on six subgroups, three for 1975 and three for 1976, to determine individual growth on the SAT given that each subgroup had taken the PSAT once and the SAT twice. Three questions were posed:

1. What was the average gain between the PSAT and the second SAT?
2. Was the average gain statistically significant?
3. Was the average gain practically significant?

The answer to each of these questions based on the statistical treatment of the data follow.

What Was the Average Gain Between the PSAT and the Second SAT?

Coaching improves SAT scores

Central to this question is the issue of whether a student can improve on the SAT if he or she gets coached. All of the coached students spent at least four hours a week for ten weeks in a SAT coaching school. The conclusion reached was affirmative.

The students who were coached made average group gains which were greater than the average gain made by the uncoached group.

Was the Average Gain Statistically Significant?

Score gains are statistically significant

The analyses of each of the groups on all of the measures produced a statistical difference at the 0.01 level for each of the three subgroups. This essentially means that it can be stated with 99 percent confidence (for each of the groups studied, coached and noncoached) that there was a difference between the PSAT and the second SAT scores for all three groups.

Was the Average Gain Practically Significant?

Practical significance prompts other questions

A comparison between a pre and postmeasure may be statistically significant; however, the practical implication or consequences of changing the conditions (e.g., a new reading program requiring texts, in-service training, etc.) may not be possible. In some cases, doing what is currently being done in an improved manner may

provide just enough change to produce no significant difference if the study were replicated. Practical significance becomes a question of a subjective value. For example, if you have $300 to $500 to spend on a coaching school, the answer would be yes; however, if you don't, the answer would be no. A related and equally important question concerns how things are working in the real world. If equal educational opportunity is in place and everyone is given the same opportunity for an education, then the answer may still be yes. On the other hand, if the way things are working systematically discriminates against certain groups, then the answer may be no. Furthermore, a basic question about the test's (SAT) predictive validity in college and life has not been answered.

Question of SAT's Predictive Validity

SAT's predictive power questioned further

There was an extensive analysis and report on the SAT recently conducted by the College Board which treated the question of the SAT's prediction capacity. In the discussion about how well the SATs predict college success, it was reported that "high school grades are still the best single predictors of college performance; but when these grades are combined with SAT scores more accurate prediction proves possible."[7] The critical point always excluded when the College Board or ETS report on the SATs predictive validity is a consideration of other variables or methods that can be used other than the SAT to improve the prediction of college success. A related issue is how much of what is being measured by the SAT is already contained in the grades given by teachers to the students.

An analysis of this concept, which was reported in NEA's *Analysis of the Wirtz Report on Declining SAT Scores* (see Appendix A for an executive summary), is repeated here to demonstrate what is meant. (In the following excerpt, HSR refers to high school record.)

> A statistic used to interpret a correlation coefficient (e.g., validity coefficient) is the *coefficient of determination*. The coefficient of determination is symbolically represented by r^2 (or correlation coefficient squared); and when the coefficient is multiplied by 100, a percentage of the variance in the two measures is determined. In the case of the HSR and first-year college grades where the validity coefficient was 0.5, the coefficient of determination (r^2) would be (0.5 × 0.5 = 0.25) 0.25. When this value (0.25) is multiplied by 100, the result is 25, or 25 percent of the explained variance. This percentage of variance (25) is interpreted to be the percentage of variance associated with the first-year college grades that can be determined by, or accounted for, in the variance of the HSR.
>
> In the 0.5 illustration, the validity coefficient of 0.5 provides the percentage of variance in college grades that is accounted for by the variance in high school grades, which was 25 or one-fourth. By using only the high school grades, 25 percent of the variance can be accounted for and 75 percent needs to be explained. Generally, this 75 percent is attributed to individual motivation and the institutional personnel who work to develop the student. This unexplained variance is the factor that makes the difference in a successful college and life experience. When the SAT Verbal and Math scores were combined with the HSR (1974), the multiple r was computed to be 0.58.
>
> The same statistical principles apply to a single or multiple correlation.
>
> Basically, a coefficient of determination may be used to determine the amount of variance explained by two or more measures correlated with the criterion. In the illustration the criterion measured is

college grades. The combined, HSR, SAT Verbal, and SAT Math were correlated with the first-year college grades and the computed multiple correlation (R) was 0.58.

The coefficient of determination (R^2) for a multiple R of 0.58 computes to be 0.34. When 0.34 is multiplied by 100, the percentage of the variance explained in the three measures (HSR, SAT Verbal, and SAT Math) is 34 percent. This leaves two-thirds of the variance in first-year college grades unaccounted for by the three measures.

Another way of viewing the increased accounted variance in absence of an established causal relationship between the measures used for prediction and the criterion measure (first-year college grades) is to evaluate how much more each measure adds to the prediction.

For example, the 1974 ETS's Validity Study Service (VSS) provided the basic data about the validity coefficients of the HSR, SAT Verbal, SAT Math, and the combined multiple correlation. These and the coefficient of determination follow.

Measure	Computed "r" between predictor (SAT scores) and criterion (college grades)	Coefficient of determination
HSR	0.50	25%
SAT Verbal	0.42	18%
SAT Math	0.39	15%
Combined	0.58	34%

HSR: Single best predictor

As can be seen from this display, the combined multiple correlation of 0.58 produces the greatest percentage of explained variance followed by HSR. However, the HSR not only provides for the most explained variance among the measures, it also accounts for much of what is being measured in the SAT Verbal and Math tests. This is revealed by the relatively small increase of variance in the combined (multiple r) coefficient of determination. The increase of 9 percent (25 percent to 34 percent) of variance accounted for in the combined measure suggests that not only is the HSR the single best predictor of first-year college grades, it also provides the same information that is tested for on the SAT. If this were not the case and if each SAT section were actually measuring something unique, there would be a greater relationship (multiple r) among the combined measures.

It appears that the validity coefficients produced in the VSS study of the 783 colleges requesting the ETS service do little to justify the use of the SAT as a predictor of first-year college grades. Further, to make a generalization about all colleges or to attempt to justify the construction of the SAT based on the relatively low reported validity coefficients raises some serious questions about whether it is the public interest or ETS's interest that is being served.

The reported information indicates that the panel answered the question about the SAT's reliability and attempted to answer the question about predictive validity.

"Unchanging standard"

The more significant validity questions about construct validity (the underlying theoretical basis of what is actually being measured by the instrument, combined with supportive statistical and logical data from research studies) and content validity (which related to the content currently being taught in the schools) were not adequately investigated or at least not reported.

To use the concept of an "unchanging standard" and to begin to investigate the changes in schools and society for 25, 20, or even five years do not suggest that the most objective approach was used to

evaluate the decline in the SAT scores. It appears that what was examined is the SAT's viability to continue in its present form as a source of ETS revenue. Furthermore, the panel served as a buffer to make this determination. There is no mention in the report about whether the SAT should be discontinued or even modified.

The CEB paid for and produced a report that did not question the validity of the use of the SAT in contemporary society. Instead, it was used as a criterion measure or an "unchanging standard." It should be noted here that the NEA requested that the CEB panel investigate the validity of the continued use of the SAT.[8]

SAT as the National Curriculum

One of the major conclusions that emerged from the Wirtz report is that the SAT was viewed as an "unchanging standard." It was portrayed as a universal truth designed to be used as a criterion to judge students' ability to succeed in college. Furthermore, the report went to great length to assure educators that, as an "unchanging standard" the SAT allows comparisons to be made with graduating seniors 5 or 20 years ago. As a constant measure, the SAT, according to ETS, ensures that "any particular score received on a current test indicates the same level of ability to do college work that the same score did 36 or 20 or 5 or 2 years ago."[9]

The "unchanging standard" challenged

The Wirtz report begged the question of how a test created in 1926 and normed and scaled in 1941 can remain *valid* through an era of curriculum changes in mathematics and science not to mention the changes in world boundaries and ideology. Furthermore, the belief that the SAT is an "unchanging standard" in a society that had evolved through the third industrial revolution (i.e., cybernetics) and three wars suggests that the motivation and desire for the "good old days" was stronger than the desire for an objective analysis of the SAT.

A national curriculum

There may be another motivation, subtle and unspoken, for defending the SAT as an unchanging standard; namely, the desire to maintain the SAT as a surrogate for a "national curriculum." It is a test that is taken by about 1.5 million students annually. Although the College Board and ETS stress that the SAT should not be used to judge school programs, teacher performance, or student progress, it is frequently used by journalists, politicians, and even educators as a quality measure of education.

The results on the coachability issue will certainly increase the number of school districts with coaching courses in the high school. This in turn will further the argument that the SAT is the first course of study that will be taught in the country and, therefore, it provides the precedent for other courses to be used in the development of a national curriculum.

Alternatives to Coaching for the SAT

There are many viable options to the SAT (and coaching for it) available for use, given the reduced number of high school graduates projected between now and 1995. It is estimated that there will be 400,000 fewer high school graduates (2.8 million to 2.4 million) between 1984 and 1987. Given this reduction in graduates and the existing available higher education classrooms, it would seem that the time has come to examine alternative selection methods and techniques. It is not necessary to continue to use an outdated and "unchanging standard" in a dynamic society with a diverse and creative youth population that deserves better than to be subjected to a 2½ hour paper and pencil test to become eligible for consideration to a college or university.

Academic Prediction Scales

Academic prediction scales

Academic prediction scales have been available for use for over 20 years. These scales are generally designed to improve the ability of high school grades to serve as predictors of college success. According to B.S. Bloom and F.R. Peters, studies have shown that with grade adjustments between high school grades and college grades correlations reach the level of +0.70 to +0.80, and in some particular schools or colleges the correlation is as high as +0.85.[6] It is not difficult to conclude that these correlations compare very favorably to aptitude and achievement test correlations, which fall in the range of 0.28 to 0.42.

With such data available for many years, the question of its lack of use has been answered with the explanation that there were too many students applying for admission. It is apparent that with declining enrollments, this explanation is inappropriate. This observation along with the evidence that the SAT discriminates against minorities, lower socioeconomic groups, and women suggest that an alternative approach should be used. The academic prediction scales described by Bloom and Peters offer at least one viable approach. This would give time to the College Board and the researchers at ETS to develop more objective and equitable measures that would treat everyone in a *just* manner.

The nation's educational objective must be to fit the desires, ambitions, and developed abilities of every student who wishes a college education to the most appropriate curriculum. This must be the approach if we are to give every child an equal educational opportunity.

Table 2. Average SAT Total Point Gain for Coached and Uncoached PSAT[1] and Two-Time SAT Takers for 1975.

	Coached Between PSAT and 1st SAT (N=65)	Coached Between 1st SAT and 2nd SAT (N=105)	Noncoached (N=455)
	(A)	(B)	(C)
Total Scores:			
PSAT	940	930	890
SAT — 1st	1,022	965	913
SAT — 2nd	1,054	1,034	934
Average gain between PSAT and 2nd SAT	114	104	44

Table 3. Average SAT Total Point Gain for Coached and Uncoached PSAT[1] and Two-Time SAT Takers for 1976.

	Coached Between PSAT and 1st SAT (N=118)	Coached Between 1st SAT and 2nd SAT (N=173)	Noncoached (N=408)
	(A)	(B)	(C)
Total Scores:			
PSAT	920	930	900
SAT — 1st	1,016	971	926
SAT — 2nd	1,063	1,065	960
Average gain between PSAT and 2nd SAT	143	135	60

[1]The PSAT is described as a shortened version of the College Board's SAT. It yields 2 scores, verbal and mathematical, on a scale of 20-80 and is directly comparable to the SAT score scale of 200-800.

Table 4. Average SAT Verbal Score Gain for Coached and Uncoached PSAT[1] and Two-Time SAT Takers for 1975.

	Coached Between PSAT and 1st SAT (N=65)	Coached Between 1st SAT and 2nd SAT (N=105)	Noncoached (N=455)
	(A)	(B)	(C)
Total Scores:			
PSAT	450	450	430
SAT — 1st	486	467	436
SAT — 2nd	505	497	447
Average gain between PSAT and 2nd SAT	55	47	17

Table 5. Average SAT Verbal Score Gain for Coached and Uncoached PSAT[1] and Two-Time SAT Takers for 1976.

	Coached Between PSAT and 1st SAT (N=118)	Coached Between 1st SAT and 2nd SAT (N=173)	Noncoached (N=408)
	(A)	(B)	(C)
Total Scores:			
PSAT	440	440	430
SAT — 1st	496	462	445
SAT — 2nd	513	505	459
Average gain between PSAT and 2nd SAT	73	65	29

[1]The PSAT is described as a shortened version of the College Board's SAT. It yields 2 scores, verbal and mathematical, on a scale of 20-80 and is directly comparable to the SAT score scale of 200-800.

Table 6. Average SAT Math Score Gain for Coached and Uncoached PSAT[1] and Two-Time SAT Takers for 1975.

	Coached Between PSAT and 1st SAT (N=65)	Coached Between 1st SAT and 2nd SAT (N=105)	Noncoached (N=455)
	(A)	(B)	(C)
Total Scores:			
PSAT	490	480	460
SAT — 1st	536	498	477
SAT — 2nd	549	537	487
Average gain between PSAT and 2nd SAT	59	57	27

Table 7. Average SAT Math Score Gain for Coached and Uncoached PSAT[1] and Two-Time SAT Takers for 1976.

	Coached Between PSAT and 1st SAT (N=118)	Coached Between 1st SAT and 2nd SAT (N=173)	Noncoached (N=408)
	(A)	(B)	(C)
Total Scores:			
PSAT	480	490	470
SAT — 1st	520	509	481
SAT — 2nd	550	560	501
Average gain between PSAT and 2nd SAT	70	70	31

[1]The PSAT is described as a shortened version of the College Board's SAT. It yields 2 scores, verbal and mathematical, on a scale of 20-80 and is directly comparable to the SAT score scale of 200-800.

Table 8. Average SAT Gain Scores for Coached and Noncoached PSAT and Two-Time SAT Takers on the Verbal, Math, and Total Scores in 1976.

GROUPS	VERBAL		MATH		TOTAL	
	PSAT	2nd SAT	PSAT	2nd SAT	PSAT	2nd SAT
Coached (A) between PSAT and 1st SAT	440	513	480	550	920	1,063
Coached (B) between 1st and 2nd SAT	440	505	490	560	930	1,065
Noncoached (C)	430	459	470	501	900	960

Table 9. Select Universities' SAT Admission Requirements for Academic Year 1980-81

University	Verbal	Math	Total
Harvard	500-800	500-800	1,000-1,600
Yale	670	680	1,350
Pennsylvania	650	660	1,310
Columbia	650	660	1,310
Pennsylvania State	450-600	450-600	900-1,200
Emory	550	600	1,150
Rutgers	490	540	1,030
George Washington	—	—	1,000
Average 1979 SAT Score for College-Bound Seniors	427	467	894

FOOTNOTES

[1]Campbell, D.T., and Stanley, J.C. *Experimental and Quasi-Experimental Designs for Research*. Chicago, Ill.: Rand McNally College Publishing Co., 1973. pp. 7-12.

[2]Educational Testing Service. *Accountability, Fairness, and Quality in Testing*. Princeton, N.J.: the Service, January 1980. p. 5.

[3]Federal Trade Commission. *The Effects of Coaching on Standardized Admission Examinations*. Washington, D.C.: the Commission, March 1979. p. i.

[4]*Ibid*., p. i.

[5]Ivens, S.H. *Backgrounder: The Effects of Coaching on a Student's SAT Scores*. New York: College Examination Board, September 1979, p. 3.

[6]*Ibid*., p. 3.

[7]College Entrance Board. *On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline*. New York: the Board, 1977. p. 9.

[8]National Education Association. *NEA's Analysis of the Wirtz Report on Declining SAT Scores*. Washington, D.C.: the Association, 1978. pp. 13-14. (Out of print).

[9]College Entrance Board. *Op. cit*., p. 8.

[10]Bloom, B.S., and Peters, F.R. *The Use of Academic Prediction Scales for Counseling and Selecting College Entrants*. Glencoe, Ill.: The Free Press of Glencoe, 1961. p. 5.

SECTION III: COMMERCIAL INVOLVEMENT IN STATEWIDE TESTING PROGRAMS

Statewide testing programs exist in nearly every state in the union, in the District of Columbia, and in Puerto Rico. There is great variation among the programs with respect to policies and procedures. There are also similarities among the programs such that they can be described according to any one of a number of features.

ETS surveys

A number of features have been surveyed and described by Educational Testing Service (ETS). In 1968 and again in 1973, ETS conducted surveys of individuals knowledgeable of the state testing programs in their respective states. The purpose of each survey was to gather information useful for constructing a profile of the state testing program. In 1968, state profiles were prepared for the areas of functions, tests, materials, and services.[1] In 1973, ETS areas of interest were program purpose, management, test population, instrumentation, data collection and processing, norms, information dissemination, and future prospects.[2]

The ETS profiles provide information useful for describing the various programs and for identifying their similarities and differences. The usefulness, however, is limited in several ways. In some respects the 1973 data are incomplete. For example, of the 33 states reporting statewide testing programs, only 28 identified the tests they used.

From a different perspective, the data are too general to suggest immediately practical uses. For example, instrumentation, i.e., tests, is of particular interest to teachers and curriculum specialists, who are probably in the best position to determine whether a match exists between what is taught and what is tested. The 1973 ETS survey focused on only four aspects of instrumentation: areas tested, tests used, whether measures had been tailored or revised for state use, and who developed tailored tests if they were used. The data were reported generally; specific tests were not reported by state, and developers of tailor-made tests were not identified.

The 1973 survey information is also dated. Because of the tremendous growth in testing during the past few years, a number of changes can be expected to have taken place during the past seven years.

NEA survey questions

For the reasons given above, NEA Research proceeded to update and extend the 1973 ETS survey and to focus primarily on the commercial aspects of instrumentation. In particular, the NEA update was designed to answer three questions:

- How many states currently conduct statewide testing programs?
- In how many statewide testing programs does commercial involvement exist?
- Which tests and test developers are involved in statewide testing programs?

Definition of Terms

For survey purposes, *statewide testing program* means one that applies to all public school districts in the state. The program may be administered through the state department of education or through a state university which may function as the testing bureau for school districts. Testing may be required of all designated students, may proceed on a random sampling basis, or may occur on a voluntary basis.

Commercial involvement means programs where test content is determined wholly or in part by publishers or by consultants whose services are purchased. *Consultant services* means assistance with test design or test development and is restricted to services involving test content such as developing test items or item pools, item validation, or test construction. The term excludes services provided by a state university where the client is the university's home state.

Survey Procedures

State department of education officials were contacted by telephone during the week of October 12-16, 1979. Individuals contacted are identified by state in Appendix A. All telephone contacts were made by one person. No interviews were recorded, and no systematic follow-up procedures were used to verify interview content. Officials in the 50 states plus the District of Columbia and Puerto Rico were included in the survey; therefore, the total number of states in the survey is reported as 52.

Results of the Survey

The states currently conducting statewide testing programs are identified in Table 1. According to the table, 9 states (17 percent of all identified states) reported having no testing program; 43 states (83 percent) reported having a statewide testing program. This represents an increase of 8 states (15 percent) with statewide testing programs since the ETS survey in 1973.

Most states report commercial involvement

Data describing the degree of commercial involvement in statewide testing programs are also reported in Table 1. For survey purposes, the degree of involvement included the categories of "None" (no commercial involvement), "Items" (use of selected test items only), "Full Test" (use of one or several complete tests), "Consultants" (use of comercial consultants) and "Full Test and Consultants." According to the data, 3 of the 43 states with statewide testing programs (7 percent) reported no commercial involvement. The remaining states with testing programs (40 or 93 percent) reported some commercial involvement. Two of the 43 states (5 percent) used only selected test items; 12 states (28 percent) used full tests only; 13 states (30 percent) used consultants only; and 13 states (30 percent) used both full tests and consultant services. Five of the 43 states (12 percent) offered but did not require the use of specific tests or test items. Four states (9 percent) involved both commercial and state consultants in their state testing programs. Because the 1973 ETS survey did not collect commercial involvement data, it is not possible at this time to say whether the 1979 data indicate increased commercial involvement in statewide testing programs.

Specific tests used and reported by state officials are identified in Table 2 by publisher, test title, and state. According to the table, McGraw-Hill is the most frequently used publisher by state programs. The company also offers the greatest

variety of tests as indicated by test title and official report. The data cannot be construed to suggest the number of students taking each test nor the local and state cost of administering the tests.

The consulting firms reported to be associated with statewide testing programs appear in Table 3. The table identifies 17 consulting agencies and the states in which services were rendered. Educational Testing Service (ETS) and National Evaluation Systems (NES) are the most active consulting firms in statewide testing programs with ETS serving 7 states and NES serving 8 states. A summary of specific consultant activity is reported by firm and state in Appendix B.

Summary and Conclusion

The purpose of this survey was to update the instrumentation focus of the 1973 ETS survey of statewide testing programs and to extend that focus to include commercial involvement. Based on the reports of 52 selected state officials, survey findings indicate that most states (83 percent) have a statewide testing program and that many of these states (77 percent) have some form of commercial involvement. States tended to use either commercially prepared tests or the assistance of consultants, although 11 states reported the use of both.

Based on these data, it is reasonable to conclude that considerable commercial involvement exists in statewide testing programs. It is also reasonable to conclude that a small number of publishers and consulting firms influence the content of tests used in the various state programs.

Table 1. 1979 Summary of State Testing Programs and Commercial Involvement

State	Testing Program				Degree of Commercial Involvement		
	No	Yes	None	Items*	Full Test Only	Consultants Only	Full Test and Consultants
Alabama		X					X
Alaska		X				X	
Arizona		X			X		
Arkansas	X						
California		X	X				
Colorado	X						
Connecticut		X					X#
Delaware		X			X		
District of Columbia		X					X
Florida		X		X			
Georgia		X					X#
Hawaii		X					X
Idaho		X				X	
Illinois		X				X	
Indiana	X						
Iowa		X			X*		
Kansas	X						
Kentucky		X			X		
Louisiana		X				X	
Maine		X				X	
Maryland		X					X
Massachusetts		X				X	
Michigan		X				X	
Minnesota		X			X*		
Mississippi		X			X		
Missouri		X					X*#
Montana		X	X				
Nebraska		X	X				
Nevada		X					X
New Hampshire		X				X	
New Jersey		X				X	
New Mexico		X			X		
New York		X			X		
North Carolina		X					X
North Dakota		X			X*		
Ohio	X						
Oklahoma	X						
Oregon		X				X	
Pennsylvania		X		X*			
Puerto Rico		X				X	
Rhode Island		X				X#	
South Carolina		X			X		
South Dakota	X						
Tennessee		X					X
Texas		X				X	
Utah		X					X
Virginia		X					X
Vermont	X						
Washington		X			X		
West Virginia		X			X		
Wisconsin		X					X
Wyoming	X						
Total	9	43	3	2	12	13	13

*Tests or items offered but not required.
#Both commercial and state university consultants are used.

Table 2. 1979 Summary of State Testing Programs by Publisher, Test, and State*

Publisher and Test	State
• American College Testing Program	
1. Adult Performance Level Exam	N.M.
• College Examination Board	
1. Degrees of Reading Power	Conn., N.Y.
2. Degrees of Writing Power	N.Y.
3. Preliminary Scholastic Aptitude Test (PSAT)	Minn.
• Harcourt, Brace, Jovanovich	
1. Metropolitan Readiness Test	D.C.
2. Metropolitan Achievement Tests	Tenn.
3. Otis Lenon Mental Ability Test	Hawaii
4. Stanford Achievement Test	Ariz., Hawaii, Nev., Tenn.*
• Houghton-Mifflin	
1. Cognitive Abilities Test	Mo.*, W.Va.
2. Iowa Test of Basic Skills	Ga., Iowa*, Md., N.D.*
3. Tests of Achievement and Proficiency	Ga.
4. Tests of Academic Progress	Mo.*
• International Business Machines (IBM)	
1. SRA Achievement Series	N.D.*, Va.
2. Iowa Tests of Educational Development	Iowa*, Minn.*
• McGraw-Hill	
1. California Achievement Test	Ala., Del., Ky. Md., Miss., N.C., Tenn.*, Wash.,
2. Career Maturity Inventory	Tenn.
3. Comprehensive Test of Basic Skills	D.C., Ky., N.M., S.C., Utah, W.Va., Wis.
4. Diagnostic Math Inventory	Ky., N.C.
5. Everyday Skills Test	D.C.
6. Prescriptive Reading Inventory	Ky., N.C.
7. Senior High Assessment of Reading Performance (SHARP)	N.C.
8. Short Form Test of Academic Aptitude	Ala., Miss.
9. Test of Performance in Computational Skills (TOPICS)	N.C.
• Psychological Corporation	
1. Differential Aptitude Test	D.C., Hawaii
2. Test of Academic Skills	Minn.*
• Teachers College Press (Columbia)	
1. Cognitive Skills Assessment Battery	S.C.

*Tests are offered but not required.

Table 3. 1979 Summary of Commercial Involvement by Consultant Firm and State

Consultant Firm	State Client
1. American College Testing Program (Washington, D.C.)	Nev.
2. American Institutes for Research (Palo Alto, Calif.)	Mich.
3. Bozler Educational Consultants (Lincoln, Neb.)	N.H.
4. Educational Testing Service (Princeton, N.J.)	Ala., Ga., Minn., Nev., N.J., P.R., Tex.
5. Institute for Behavioral Research and Creativity (Salt Lake City, Utah)	Utah
6. Instructional Objectives Exchange (Los Angeles, Calif.)	Va.
7. Intran (Minneapolis, Minn.)	La.
8. McGraw-Hill (New York, N.Y.)	D.C.
9. National Evaluation Systems (Amherst, Mass.)	Conn., Ga., Hawaii, Md., Mass., N.J., R.I., Va.
10. National Testing Service (Durham, N.C.)	Del., La.
11. Northwest Evaluation Association	Wisc.
12. Northwest Regional Laboratory (Portland, Oreg.)	Alaska, Idaho, Oreg.
13. Research Management Corporation (Portsmouth, N.H.)	N.H.
14. Research Triangle (Raleigh, N.C.)	Ill., Maine
15. Scholastic Testing Service (Bensenville, Ill.)	N.C., Tenn., Va.
16. Science Research Associates, International Business Machines (Chicago, Ill.)	Mo.
17. Touchstone Applied Science Associates (Elmsford, N.Y.)	Conn., N.Y.

FOOTNOTES

[1]Educational Testing Service. "State Testing Programs: A Survey of Functions, Tests, Materials, and Services," Princeton, N.J.: The Service, 1968. (TM 003 001)

[2]Educational Testing Service. "State Testing Programs: 1973 Revision." Princeton, N.J.: The Service, 1973. (TM 003 397).

SECTION IV: NEA POSITION ON TESTING

Historical Background

In February 1973 the National Education Association Center for Human Relations held a national conference in Washington, D.C. The theme of the three-day conference was "Tests and Use of Tests—Violations of Human and Civil Rights." The objectives of the conference were:

- To examine current attitudes about the educational value of standardized tests, especially as they affect the culturally different learner.
- To explore alternative measurement and evaluation processes that would be helpful tools in the education process.
- To create greater national awareness of the need for concerted action to prohibit the use of test scores as indicators of growth potential, especially for the culturally different learner.

NEA moratorium

One major outcome of the conference was a recommendation to the NEA Representative Assembly concerning standardized tests. Meeting that summer, the Assembly adopted Resolution 72-44 on "Standardized Testing." The Resolution encouraged the elimination of standardized group tests of intelligence, aptitude, and achievement until a critical appraisal, review, and revision of current testing programs had been conducted.[1] Known as the NEA moratorium on testing, the Resolution remained in effect until 1978 when it was revised.

Several events during the years following the moratorium suggested a need to reexamine NEA testing policy. The 1972 moratorium had successfully alerted the public to the dissatisfaction of many educators with existing tests and testing practices. The dissatisfaction, however, needed elaboration, especially in light of widespread and often uncritical acceptance of standardized test scores, misinterpreted test results, and increased demand for testing. There was, too, professional recognition that some tests, if carefully constructed, could be instructionally useful.

On Further Examination and NEA Response

One event in particular precipitated an Association response. The event was the release in 1977 of *On Further Examination*, published by the College Examination Board (CEB). The study was conducted by an advisory panel of 21 people chaired by former Secretary of Labor Willard Wirtz. The purpose of the study was to investigate declining Scholastic Aptitude Test (SAT) scores among high school students. The study was sponsored and funded by CEB and Educational Testing Service (ETS). CEB sponsors the SAT; ETS develops and administers it.

On Further Examination was the study of the 14-year decline in SAT scores from 1963 to 1977. Verbal scores had dropped 49 points, from 478 in 1963 to 429 in 1977. During this same period, mathematics scores had dropped 32 points, from 502 to 470. The question posed to the Wirtz panel was, why?

The report represented a comprehensive analysis of social and educational change believed to be reflected in test scores. The change and resulting test score decline were proposed to have occurred in two stages, each characterized by different causal factors. The first premise explained score changes from 1963 to 1969 as the result of a changing population of students taking the SAT. The population purportedly included "larger proportions of students of characteristically lower-scoring groups of students."[2] The second premise attributed the decline from 1970 to 1977 to various social and educational factors.[3]

NEA response: SAT not examined

NEA prepared three responses to the Wirtz report. The first response appeared as an editorial by NEA President John Ryor in the November-December 1977 issue of *Today's Education*. Ryor acknowledged in the editorial the panel's effort to be fair, to demonstrate some understanding of the different tasks of teachers, and to express awareness of some of the criteria guiding the use of SAT scores. Ryor concluded, however, that the report could provoke a misuse of test data by legislatures who would see only declining test scores and ignore the caution against imposing upon the schools more rigidity and uniformity.[4] Ryor's primary objection was that the panel examined test results, not the test itself. Thus, the fundamental and unanswered question was: "Should a SAT test which hasn't changed significantly in 36 years be allowed to become a major determinant of school curriculum?"[5]

NEA response: unfinished study

The second response was a booklet entitled *On Further Examination of 'On Further Examination'* published in 1977 by NEA Instruction and Professional Development. This publication commended the Wirtz panel for its lack of indictment, attention to multiple rather than single questions and theories, and consultation with some experts. Nevertheless, the publication argued that the examination of declining SAT scores was unfinished. The paper analyzed panel comments about the SAT, teaching, and selected aspects of society. The paper noted that panel members carefully avoided an analysis of the test itself and that further examination of the declining test scores should address many more issues such as questions of validity (particularly predictive validity) and cultural bias, the assumption that educational content and performance standards remain unchanged over time, and whether SAT could or should measure such skills as thoughtful and critical reading and careful writing.[6]

The third response was prepared by the NEA Special Committee on Declining SAT Scores. Appointed in 1977 by John Ryor, the five-member committee was charged with three tasks:

- To analyze *On Further Examination*
- To review NEA's current policy on testing
- To develop a set of policy recommendations.

NEA response: technical considerations

In 1978 the NEA committee submitted the results of its investigation entitled *NEA's Analysis of the Wirtz Report on Declining SAT Scores*. (A copy of the executive summary of this report appears in Appendix C.) Among conclusions reached were:

- The conclusions in the Wirtz report exceed those that can be reasonably drawn from the provided descriptive statistics.[7]
- The SAT has been constructed to ensure test reliability at the expense of test validity.[8]
- Item selection is based more on the power of items to differentiate among students than on the match of items with instructional content.[9]
- The value of the SAT is questionable when questions of validity are addressed.[10]

- The two premises used to explain the SAT score decline lack objective documentation and are not generalizable.[11]
- There is no evidence to support the view that students learn less today than their counterparts in the past.[12]

NEA testing policy reviewed

The remaining charges of the NEA Special Committee were to review existing testing policy and prepare policy recommendations. Based on its review of existing policy, the Committee concluded that several policy changes were desirable. The Committee believed that some tests could be instructionally useful to teachers and that such tests should be supported. The Committee also believed that many tests were inappropriate for educational measurement and evaluation and that steps should be taken to help teachers become better informed about the meaning of tests and the use of test data.[13]

NEA policy revised

Recommendations of the Special Committee were presented to members of the 1978 Representative Assembly meeting in Dallas, Texas. In response to the recommendations, the Assembly revised the 1972 Resolution. The new Resolution 78-82 on "Standardized Testing" recognized that student testing could serve important educational purposes such as diagnosing learning needs, prescribing instructional activities, and measuring student progress in curriculum content.[14] The Resolution supported the use of tests prepared or selected by the teacher and made explicit NEA opposition to standardized tests which are:

- Damaging to a student's self-concept and contribute to the self-fulfilling prophecy whereby a student's achievement tends to fulfill the negative expectations held by others.
- Biased against those who are economically disadvantaged or who are culturally and linguistically different.
- Used for tracking students.
- Invalid, unreliable, out-of-date, and restricted to the measurement of cognitive skills.
- Used as a basis for the allocation of federal, state, or local funds.
- Used by book publishers and testing companies to promote their financial interests rather than to improve measurement and instruction.
- Used by the media as a basis for invidious public comparisons of student achievement test scores.
- Used to test performance levels as a criterion for high school graduation.
- Used to evaluate teachers.

1980 NEA policy

The Wirtz report was an occasion for the Association to explain why it opposed so many tests and testing practices. Recent proposals for truth-in-testing legislation have provided opportunities to describe testing changes the Association advocates. (See Appendix D for the 1980 Resolutions which further elaborate the NEA position on testing.)

Two Testing Changes NEA Advocates

On August 1, 1979, and again on October 10, 1979, the NEA testified before the Subcommittee on Elementary, Secondary, and Vocational Education. The Committee was considering two proposals for truth-in-testing legislation. Both proposals, whose contents are discussed in Section V, concerned truth and disclosure legislation. During testimony the Association expressed the belief that certain

changes in testing could improve tests and the way they are used. A copy of the NEA analysis of the federal proposals appears in Appendix E. The changes, elaborated for this publication, are discussed below.

Criterion-Referenced Tests

One change already occurring but on a limited scale is the use of criterion-referenced tests. Criterion-referenced tests are those in which individual performance is described in terms of specific instructional content or performance objectives rather than in terms of the performance of others.

The need for criterion-referenced tests

The popular notion of criterion-referenced testing may have come from the distinction made by Robert Glaser in 1963. Concerned about the failure to use tests for instructional purposes, Glaser appealed for tests that could be interpreted directly in terms of defined educational content. He distinguished between test scores whose interpretation indicated what an individual could actually do and test scores whose interpretation indicated what an individual could do when compared to others. The former were criterion-referenced tests; the latter, norm-referenced.[15]

Criterion-referenced tests are an alternative approach to traditional tests. They have several characteristics which make them instructionally useful. If well designed and carefully constructed, criterion measures describe with considerable clarity the specific knowledge and skill measured. Thus, teachers can select or develop tests better matched with actual instruction and educational objectives. The measures will be more accurate, the quality of test data will be improved, and the information about achievement and progress will better serve the goals of improved instruction.

Criterion-referenced tests are designed to describe performance relative to instructional content. Measures that succeed in this respect can be expected to make more sense to students and teachers. Success and error can be more readily understood in terms of specifics rather than vague abstractions. The interpretation of test scores in terms of specific content and skill also makes more manageable the task of understanding error and correcting it.

Criterion testing will not end the current practice of norm referencing. The belief has somehow emerged that some tests are criterion referenced and others are norm referenced but that a test cannot be both. In fact, a test score can be interpreted both ways provided test content is precisely described.

The problem of clarity

Nor will criterion testing be problem free. One problem is the difficulty of achieving descriptive clarity of the content and behaviors to be measured. Various frameworks have been developed to help promote descriptive clarity. Among available frameworks are various theoretical constructs such as cognitive and affective domains and structure-of-intellect models; instructional, behavioral, or performance objectives; content-processing matrices; and formal rules for item development.

The problem of meaning

A second problem is the difficulty of attributing meaning to test scores. Criterion scores have been expressed as expectancies, predictors; diagnostic signs, and indicators of mastery. The terms imply a performance standard or cutoff point. G.V. Glass has argued that existing methods of determining criterion scores are arbitrary and that interpretations based on absolute standards are meaningless given existing knowledge.[16] Glass asserts that "the only sensible interpretation of data from assessment programs will be based solely on whether the rate of performance goes

up or down."[17] If this is the case, then new interpretive guidelines will be necessary if indicators of the direction and rate of change are to make instructional sense.

Buros Reform Proposal

Emphasis on measurement

Criterion testing is one change encouraged by the NEA. A second change that holds promise for educational measurement is embedded in Oscar Buros' proposal for test reform. Buros favored tests built for the purpose of measurement rather than differentiation.[18] To achieve this end, he proposed developing different tests to measure the achievement of groups and the achievement of individuals. Group tests would be used to measure groups such as schools or school systems with common objectives and learning environments. Individual tests would be used to measure individuals. Group tests could cover both common objectives and objectives unique to a school or school system and could be administered to a sample of students. Individual tests would cover those objectives unique to local objectives for which measures of specific individual growth would be desired.

Emphasis on usefulness

Buros believed methods of reporting test data could be simplified. He advocated local rather than national norms, raw score means, and frequency distributions of raw score means calculated for item scores and total test scores. He also believed that individual scores could be more meaningfully reported if the raw score were reported as a percentage of the possible total score and also if percentile rank within grade were reported.[19] For example, a descriptive record of 80/65 for a given student would indicate that within a given grade that student successfully answered 80 percent of the items and scored as well or better than 65 percent of the other students locally.

Example of Feasibility of Some Advocated Changes

The Association favors the use of criterion measures for both group and individual tests, and it favors reporting data in more usable forms. The feasibility of accomplishing this for groups on a large-scale basis has already been demonstrated by the National Assessment of Education Progress (NAEP). NAEP is an example of criterion-referenced testing. The broad purpose of NAEP is to measure the nation's educational progress, and the function of the various test exercises is to describe achievement in terms of educational content and specific instructional and behavioral goals. Exercises are statistically sorted into booklets, booklets are administered to individuals selected to represent significant characteristics such as age and geographical region, and test data are reported by subject area, age group, and instructional content.

There are many features that distinguish NAEP testing from standardized achievement tests. Test exercises are developed to measure educational objectives consistent with instruction. The selection of test items is based on their match with instructional content rather than on their power to discriminate among students. Sampling procedures allow for the assessment of many cognitive and affective objectives without subjecting students to lengthy test sessions. Results of the data are also easily understood by professional and lay audiences.

There is an additional feature of the NAEP program worth noting. NAEP is governed by a relatively open testing policy. That is, the theoretical and practical aspects of test development are richly documented and accessible. Furthermore, reported data are accompanied by actual test items and their correct answers (up to half of all NAEP items are released after test administration). Thus, one knows the

objective measured, the instrument of measure, and the results which are reported by objective and item and are portrayed in various ways. This disclosure has responsibly informed people about the test, and it has also provided educators with information and ideas that are instructionally useful. (See Appendix F for NEA's letter of support to the Education Commission of the states regarding the National Assessment of Education Progress.)

Other Changes NEA Supports

NEA supports open tests

NEA supports the idea of open tests and believes that the release of all test items and their answers will be a significant change in educational measurement. The Association respects the idea of secure test items prior to test administration provided there is reason to believe they are well designed, well constructed, theoretically sound, and instructionally relevant. After test administration, NEA believes students have a right to inspect their own performance and to have the opportunity to learn from their successes and errors.

The Association supports a number of other efforts to improve testing in the United States. Among such efforts are local test development; construction of a variety of measures including observation and student self-reports; sequential testing where items and tests are tailored for individuals; item banks with items classified, stored, and retrieved according to specified content, format, and difficulty; and computer-generated tests constructed to meet certain specifications.

For over a decade NEA has advocated change in the way testing is viewed and practiced in the United States. The Association believes that change will constructively occur when testing and instruction aim toward the same objectives and are designed for the same purpose of providing the best education possible for all individuals.

FOOTNOTES

[1]National Education Association. *1973-74 NEA Handbook*. Washington, D.C.: the Association, 1973, p. 81.

[2]College Examination Board. *On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline*. New York: the Board, 1977, p. 45.

[3]*Ibid.*, p. 45.

[4]Ryor, John. "Declining SAT Scores." *Today's Education* 66:6; November-December 1977.

[5]*Ibid.*, p. 8.

[6]National Education Association. *On Further Examination of "On Further Examination."* Washington, D.C.: the Association, 1977. pp. 9-11.

[7]National Education Association. *NEA's Analysis of the Wirtz Report on Declining SAT Scores*. Washington, D.C.: the Association, 1978. p. 8. (Out of print)

[8]*Ibid.*, p. 9.

[9]*Ibid.*, p. 10.

[10]*Ibid.*, p. 11.

[11]*Ibid.*, p. 25.

[12]*Ibid.*, p. 25.

[13]*Ibid.*, pp. 54-55.

[14]National Education Association. *1978-79 NEA Handbook*. Washington, D.C.: the Association, 1978. p. 239.

[15]Glaser, Robert. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." *American Psychologist* 18:519-521; August 1963.

[16]Glass, G.V. *Standards and Criteria*. Occasional Paper No. 10. Kalamazoo, Mich.: Evaluation Center, Western Michigan University, 1977. pp. 14-41.

[17]*Ibid.*, p. 45

[18]Buros, Oscar K. "Fifty Years in Testing: Some Reminiscences, Criticisms, and Suggestions." *Educational Researcher* 6:15; July-August 1977.

[19]*Ibid.*, pp. 14-15.

RECOMMENDED READING

Buros, Oscar K. "Fifty Years in Testing: Some Reminiscences, Criticisms, and Suggestions." *Educational Researcher* 6:9-15; July-August 1977.

Glass, G.V. *Standards and Criteria*. Occasional Paper No. 10. Kalamazoo, Mich.: Evaluation Center, Western Michigan University, 1977. 49 pp.

Nairn, Allan et al. *The Reign of ETS: The Corporation That Makes Up Minds*. The Ralph Nader Report on the Educational Testing Service. Washington, D.C.: Ralph Nader Organization, 1980. 554 pp.

National Assessment of Educational Progress. *Changes in Mathematical Achievement, 1973-78*. Report No. 09-MA-01. Denver, Colo.: Education Commission of the States, August 1979. 31 pp.

National Education Association. *NEA's Analysis of the Wirtz Report on Declining SAT Scores*. Washington, D.C.: the Association, 1978. 57 pp. (Out of print).

College Examination Board. *On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline*. New York: the Board, 1977. 75 pp.

SECTION V: TRUTH-IN-TESTING LEGISLATION

California legislation

In 1978 the California legislature examined a proposal concerning information about tests. The proposal required test publishers to disclose to the California Postsecondary Education Commission descriptive information about test content, test validity, standards, administration, expenses incurred, and income. It also required publishers to provide to test takers descriptive information about test content, test purpose and use, treatment of scores, and score ownership. The proposal applied only to standardized tests administered to 3,000 or more students for the purpose of postsecondary admissions selection. The legislation was enacted in September 1978 and became the first truth-in-testing law in the United States.

New York legislation

In 1979 similar legislation was enacted in New York. The New York law applied to tests used for postsecondary and professional school admission selection and specifically excluded civil service exams and tests used for other purposes. It required the disclosure of similar kinds of descriptive information required in California. Unlike California, which required disclosure only of test questions equivalent to those actually used, the New York law required full disclosure of test items actually used. It was the full disclosure clause which made the New York legislation controversial, even after it was enacted in July 1979.

Similar proposals in other states—Florida, Ohio, and Pennsylvania, for example—and two at the federal level concern truth-in-testing. None of these proposals has yet been enacted, but others will undoubtedly be proposed and eventually made into law as the movement gains momentum.

Current truth-in-testing proposals and laws are aimed at standardized tests and represent notice and disclosure legislation. As they are currently conceptualized, the proposals have been viewed as a variation of consumer protection legislation. The legislation recognizes the right of consumers to be informed about the products and services they purchase. Consumers of testing include students who are tested and who often pay test fees, educational organizations such as the American Medical Association for whom special tests are developed, and the states with constitutional responsibility for public education.

A number of issues are involved in truth-in-testing debates. Some of the issues, although important, do not address the legislation directly. Some of these issues, identified in *Searching for the Truth about Truth-in-Testing Legislation* published by the Education Commission of the States, involve undifferentiated discussion of tests, undifferentiated discussion of the information needs of various individuals and groups in education, and narrow focus on certain kinds of test performance.[1] These are issues where testing opponents and proponents tend to talk past each other rather than tackle the issue directly. As already mentioned, these are issues tangential to most legislative proposals.

Primary issues

Issues that tend to bear directly on the legislation revolve around five major subjects: the need for tests, test publishers, test quality, the need for testing legislation, and the consequences of testing legislation. The arguments on both sides of each issue are summarized below.

The Need for Tests

Selection decisions

The kinds of tests under consideration are measures of achievement and aptitude. Their use is restricted usually to postsecondary admissions selection. Proponents of current practices point to the large numbers of students attempting to gain admission to postsecondary schools and the need for information for selection purposes. With limited budgets, space, and curricular programs, institutions need information to help them select those students best qualified and most likely to complete successfully a course of study. Test scores can supply this information more objectively than can other information sources. It is also argued that tests can help students self-select postsecondary schools consistent with their own abilities and preferences.

Opponents argue that with college enrollments dropping and universities in need of revenue, the need to select and reject certain students has diminished. The tests systematically penalize certain groups of students and function more effectively as instruments to maintain the status quo. The test results adversely affect the educational aspirations and employment opportunities of many individuals and should not be used any way in publicly supported institutions or in institutions that compete for and accept federal tax dollars.

Test Publishers

The test publishers in question are those who produce standardized tests used for postsecondary admissions selection. The exemplar chosen is often Educational Testing Service (ETS), producer of one of the more common tests, the SAT, used for selection purposes. The central issue here is responsibility, or accountability, as it is called in the public sphere.

Public and market accountability

Opponents of testing legislation argue that test publishers are accountable. In the marketplace they are one of numerous competitive testing companies with comparable financial resources. Therefore, they cannot be regarded as a monopoly. Test publishers make their products and services available to institutions which are not forced to use them but rather exercise free choice in test selection and test use. Test publishers are accountable to clients who design testing programs and request special tests for program purposes. Test publishers are also accountable to the public under whose laws they are regulated and whose educational members have access to many reports prepared regularly for their benefit.

Proponents of the legislation argue that publishers are in the business of measuring minds and so exercise great influence over what to think and how to think. Claims made for the power of the tests and for the science of their measurement lack convincing evidence, but the claims are nevertheless repeatedly made. The largest number of people who actually purchase and "use" tests and testing services are students who have inadequate knowledge of the nature of the measure to which they submit and the use that will be made of the data they individually provide.

Test Quality

Opponents of testing legislation argue that tests are theoretically and technically sound, given existing knowledge, and reflect social and educational values associated with intellectual development and cognitive power. Opponents do not claim that tests are designed for comprehensive personal, social, or intellectual measurement; nor do they claim that existing tests can assess such qualities as creativity,

imagination, and persistence. Specific item weaknesses have been acknowledged but often with the defense that items undergo an extensive review and revision process and that efforts are made constantly to improve test content. Technically speaking, opponents agree that test validity is difficult to achieve, but opponents argue that efforts are made continually to gather validity data. They also argue that tests do what they were designed to do, and nothing more. They were not designed to predict with perfect accuracy the future of individuals. They were designed rather to improve short-term predictions about people; and this, it is claimed, they generally do.

Conceptual and technical arguments

Proponents of testing legislation challenge current theoretical models of intelligence or innate capacity. What the tests measure, they say, are skills and content that can be taught. The use of these tests consequently influences what is taught, what is learned, and what is thought. Tests also fail to capture the range of human qualities that are involved in various human endeavors such as pursuing a course of study and working toward an academic degree. Technical arguments by proponents frequently involve criticism of specific test items as a way to illustrate a range of problems with the test such as cultural bias, ambiguity, and over-simplified logic. Technical quality is challenged particularly with respect to predictive validity which opponents of the tests say lacks convincing evidence and does not improve upon existing predictors such as grade point average.

The Need for Testing Legislation

Opponents of testing legislation argue that the need for legislation has not been demonstrated. They reject the logic behind arguments that test producers and tests control or adversely influence educational content and ways of thinking. They refute arguments for more information by noting the amount and kind of information already provided and make the case for secure testing in the name of quality control. Government regulation, they argue, is unnecessary and in the case of federal regulation violates states' rights to control education. They also argue that such regulation is obtrusive and unconstitutional intervention.

Proponents of testing legislation argue that more information about tests is necessary if tests are to be wisely chosen and judiciously used. This information can be supplied by test publishers who have steadfastly refused to release it. They argue that institutions are bound by various state and federal regulations that require them to meet certain standards and achieve certain aims. The federal government has been involved in education since military academies were established in the eighteenth century but particularly in the post World War II period. They also argue from an analogy between test materials and services and consumer products now under federal regulation. They argue that the consumer has a right to be fully informed about the nature of the product or service he or she purchases whether the product is a hair dryer, automobile, or test.

The question of information control

Consequences of Legislation

The consequences of testing legislation are, from the proponents' point of view, largely positive. Legislation will force producers to be accountable to test users and test takers, will result in the dissemination of quality information, will open tests to the scrutiny of many people including professional educators and researchers, and will result ultimately in improved tests and improved use of test information.

The consequences of testing legislation are viewed less optimistically by opponents. Test producers argue that proposed changes, particularly full disclosure clauses, will increase the cost of test production. These costs in turn will be passed on to students who will pay more for each test; poor students will be affected most. Test producers believe testing legislation will adversely affect test quality and will lead to the withdrawal of some tests in states with testing controls. Ultimately decision makers will be forced to rely on less accurate information and, therefore, to make arbitrary decisions about individuals.

Open versus Secure Testing

Full disclosure

By far the most explosive issue in truth-in-testing legislation is the full disclosure clause which mandates the release of actual test questions and answers soon after the test has been given. Open testing means test disclosure. Secure testing means no test disclosure even after the test has been administered. The issue of open versus secure testing involves test information and its accessibility.

Disclosure clauses in truth-in-testing legislation would open tests after administration to public scrutiny. Arguments in support of open testing appeal primarily to the test taker's right to be informed and the test producer's obligation to provide that information.

The right to be informed is sometimes treated as a right in itself or as a matter of ethics.[2] When an individual is tested, it is argued that he or she has the right to know the results of the examination, the meaning attributed to the results, and the original data. Usually the discussion of rights shifts to decision making where test results are involved in decisions such as college admission that affect the test taker. With more at stake, the test taker has the right to examine and judge the kind of test data he or she provides for decision-making purposes. Most often the right-to-know argument is expressed as a matter of fairness where personal feelings are set aside in an effort to achieve a balance of conflicting interest. If the test taker must submit to testing and accept the results, then fairness involves the opportunity to be fully informed of the data and the standards of judgment.

Because tests function as instruments of social policy, test producers have a responsibility to inform test takers and the public about the instruments provided. This is an accountability argument, and it is appropriate in the public sector. This argument affirms the belief that those entrusted with public institutions must be accountable to the public which supports them. One aspect of this accountability is to increase information about the instruments used to decide who will and who will not attend public institutions.

The case for technical quality

Arguments to support secure testing involve test quality, controlled costs, and constitutional questions. Secure testing is believed to be a necessary condition for test quality. One technical characteristic of quality is test validity. Test validation is a process of providing evidence that the test measures what it is supposed to measure. In cases where multiple forms of each test are developed each year and for successive years, some effort must be made to make the various test forms within a given year and across successive years equally valid. The procedure for establishing the equality of multiple tests across time is called equating. It involves reusing test items in successive test administration. Open testing would require that test content be disclosed sometime after test administration. This disclosure would damage test validity by exposing those items intended for reuse. Thus, it would end current equating practices. Other concerns with test validity involve those subject areas for which a limited number of test items exist. Open testing would eventually expose all

items and would increasingly erode test validity. The end result would be a diminished confidence in tests whose quality and usefulness would be eroded through exposure.

The problem of cost

Open testing requires that new test forms be continually developed for each test administration and that new methods of equating the forms be developed. The process of research and development needed to achieve this would be expensive. These costs would eventually be passed on to test takers. Thus, the legislation would force costs upward and would affect everyone. State and federal regulation would inflate the cost of testing.

Open testing is also viewed as unconstitutional. One claim is that open testing infringes on First Amendment rights interpreted in this argument as an institutional right to decide who will be admitted to college and also as an individual researcher's right to determine whether or not her or his research will be made public. The latter enters into the debate because some research on testing is conducted by private individuals who have no financial relationship with test publishers but whose research helps establish various technical qualities of tests.

Constitutional considerations

A second claim invokes the Fifth and Fourteenth Amendments. The Fifth Amendment prohibits the federal government from depriving any person of life, liberty, or property without due process of law. The Fourteenth Amendment extends the provisions of the Fifth Amendment to include state governments. Private property in testing legislation refers to tests and related test data. The claim of private property is strengthened by test copyrights which bring tests under the protection of the Federal Copyright Act of 1976.[3] Given existing law, the disclosure clause would deprive test producers of exclusive rights to their tests and would in effect destroy their value for future use.

The issue of open versus closed testing is complex. It has attracted considerable attention from various groups and individuals, and it is likely to persist. For those interested in following the debate nationally and within their respective states, two well-documented and reasoned publications are worth study. One paper is *The Debate Over Open Versus Secure Testing: A Critical Review* written by Andrew Strenio, Jr.[4] Prepared for the National Consortium on Testing, the paper examines the case for testing legislation, the case for perpetuating existing test practices, and the strengths and weaknesses of the arguments. The second publication was prepared by the Education Commission of the States and is entitled *Searching for the Truth about Truth-in-Testing Legislation*.[5] Prepared for legislators, the report pays close attention to legislative arguments, existing law, and legal implications of testing legislation.

The NEA Position on Truth-in-Testing Legislation

NEA supports truth-in-testing legislation

In June 1979 the NEA Representative Assembly voted to urge a congressional investigation of the standardized testing industry, the tax-exempt status of testing companies, and the need for truth-in-testing legislation. In August and again in October 1979, the Association presented testimony on two federal truth-in-testing legislation proposals being studied by the Subcommittee on Elementary, Secondary, and Vocational Education. The proposals, the Truth-in-Testing Act of 1979 (H.R. 3564) and the Educational Testing Act of 1979 (H.R. 4949), both represented notice and disclosure legislation which the Association supported. (See Appendix E.)

The Association favors truth-in-testing legislation. The legislation represents an effort to promote public accountability of the testing industry and also of the schools. The legislation will make possible access to information necessary for responsible test selection and use. The legislation will further the aim of needed test reform. Above all, truth-in-testing legislation will provide information to the people who can benefit most from open testing and full disclosure: students whose intellectual growth and development can be enhanced by personal knowledge of their measured achievement and whose preparation for college and career entry can benefit from quality test data timely provided.

FOOTNOTES

[1]Education Commission of the States. *Searching for the Truth About "Truth-in-Testing" Legislation.* Report No. 132. Denver, Colo.: the Commission, 1980. pp. 12-14.

[2]Strenio, Andrew, Jr. *The Debate Over Open Versus Secure Testing: A Critical Review.* Staff Circular No. 6., Cambridge, Mass.: The Huron Institute, 1979. p. 6.

[3]Education Commission of the States. *Op. cit.*, p. 38.

[4]Strenio, Andrew, Jr. *Op. cit.* pp. 1-72.

[5]Education Commission of the States. *Op. cit.* pp. 1-89.

RECOMMENDED READING

Education Commission of the States. *Searching for the Truth About "Truth-in-Testing" Legislation.* Report No. 132, Denver, Colo.: the Commission, 1980. 89 pp.

Strenio, Andrew, Jr. *The Debate Over Open Versus Secure Testing: A Critical Review.* Staff Circular No. 6., Cambridge, Mass.: The Huron Institute, 1979. 71 pp.

Appendix A

SURVEY PARTICIPANTS FOR COMMERCIAL INVOLVEMENT IN STATEWIDE TESTING PROGRAMS

1979 Survey Participants

I. *OFFICIALS OF STATE BOARDS OF EDUCATION*

Alabama:	Clinton Owens	(205) 832-3402	Missouri:	Charles Foster	(314) 751-3545
Alaska:	Ernest Polley	(907) 465-2967	Montana:	Bill Connett	(406) 449-3693
Arizona:	Steve Stevens	(602) 255-5837	Nebraska:	Harriet Egerson	(402) 471-2444
Arkansas:	James Washburn		Nevada:	George Barnes	(702) 885-5700
	Connie Darden	(501) 371-1464	New Hampshire:	James Carr	(603) 271-3740
California:	Dale Carlson	(916) 445-4338	New Jersey:	Carl Johnson	(609) 292-4450
Colorado:	James Hennes	(303) 839-2111	New Mexico:	Bayla Nochumson	(505) 827-2282
Connecticut:	Douglas Rendone	(203) 566-8250	New York:	Windsor Lott	(518) 474-5099
	George Kinkaide	(203) 566-7232	North Carolina:	Robert Evans	(919) 733-3813
Delaware:	Robert Bigelow	(302) 678-4583	North Dakota:	Hank Landes	(701) 224-2391
District:	Robert Farr	(202) 724-4164	Ohio:	Ken Higgins	(614) 466-4868
Florida:	Thomas Fisher	(904) 488-8198	Oklahoma:	James Casey	(405) 521-3196
Georgia:	Elizabeth Creech	(404) 656-2661	Oregon:	Susan Holmes	(503) 378-3583
Hawaii:	Selvin Chin-Chance	(808) 656-2661	Pennsylvania:	Robert Coldiron	(717) 787-4234
Idaho:	Karen Underwood	(208) 384-2113	Puerto Rico:	Edith Vasquez	(809) 754-0964
Illinois:	John Alford	(217) 782-4984	Rhode Island:	Martha Highsmith	(401) 277-3126
Indiana:	Ronald Hartman	(317) 927-0241	South Carolina:	Terry Helsley	(803) 758-8610
Iowa:	Max Morrison	(515) 281-5274	South Dakota:	Robert Huckins	(605) 773-3371
Kansas:	Judy Hamilton	(913) 296-3201	Tennessee:	Jesse Warren	(615) 741-1099
Kentucky:	Armand Discontini	(502) 564-4394	Texas:	Keith Cruse	(512) 475-2066
Louisiana:	Hugh Peck	(504) 342-3750	Utah:	Dave Nelson	(801) 533-5461
Maine:	Betty McLaughlin	(207) 289-2033	Vermont:	Karlene Russell	(802) 828-3111
Maryland:	William Grant	(301) 796-8300 Ext 328	Virginia:	Richard Boyer	(804) 786-2624
Massachusetts:	Mathew Towle	(617) 727-0190	Washington:	Gordon Ensign	(206) 753-3449
Michigan:	Edward Roeber	(517) 373-8393	West Virginia:	Doris White	(304) 348-3230
Minnesota:	William McMillan	(612) 296-6002	Wisconsin:	James Gold	(608) 266-3390
Mississippi:	Rex Pouncey	(601) 354-6979	Wyoming:	Lynn Simons	(307) 777-7673

II. *OTHER PARTIES CONTACTED*

1. EDUCATION COMMISSION OF THE STATES: Jack Schmidt (303) 861-4917
2. SCHOLASTIC TESTING SERVICE: John Kauffman (313) 665-0089
3. TOUCHSTONE APPLIED SCIENCE ASSOCIATES: Dr. Bertram Koslin (914) 592-2630

Appendix B

SUMMARY OF CONSULTANT ACTIVITY BY FIRM AND STATE

1. AMERICAN COLLEGE TESTING PROGRAM

Nevada: ACT assisted in establishing the validity of items for the Nevada Competency Test Program. This test is currently given in the ninth grade and eventually will be given in the twelfth grade.

2. AMERICAN INSTITUTES FOR RESEARCH

Michigan: AIR assisted in developing the tests used in the Michigan Educational Assessment Program. Under the program tests are now given in grades 4, 7, and 10.

3. BOZLER EDUCATIONAL CONSULTANTS

New Hampshire: BEC is assisting in field testing and report writing for the New Hampshire Educational Assessment Program.

4. EDUCATIONAL TESTING SERVICE

Alabama: ETS consulted on the validity of a state competency test piloted in 1979. The test will eventually be given in grades 3, 6, and 9.

Georgia: ETS advised on the development of the Georgia Criterion Reference Tests. These include tests in reading, mathematics and career development in grades 4, 6, and 8 and a tenth grade test in mathematics and communications skills. The current contractor is the University of Georgia.

Minnesota: ETS developed the PSAT and the SAT tests, which are offered to school districts through the University of Minnesota's Student Counseling Bureau.

Nevada: ETS advised on the procedure for writing test items used in the Nevada Competency Test Program. This test is currently given in the ninth grade and will eventually be given in the twelfth grade.

New Jersey: ETS assisted in item development for the New Jersey Minimum Basic Skills Tests. These competency tests in reading and mathematics are given in grades 3, 6, 9, and 11. The current contractor for new items is NES.

Puerto Rico: ETS consults on a continuing basis regarding the validation and interpretation of results for Pruebas de Stresas Basicas (Tests of Basic Skills). These include achievement tests in mathematics and Spanish reading in grades 2 and 3, plus tests in English given in grades 4, 5, and 9. ETS plays a similar role with respect to the Prueba de Abilidad General, which is given in grades 4, 7, and 10.

Texas: ETS is currently developing an item pool for the Texas Assessment of Basic Skills. These tests cover math and reading and are administered in grades 5 and 9.

5. INSTITUTE FOR BEHAVIORAL RESEARCH AND CREATIVITY

Utah: IBRC advised on goal development and item validity for sections of the Utah Statewide Assessment Battery, which deal with emotional maturity, music, and art.

6. INSTRUCTIONAL OBJECTIVES EXCHANGE

Virginia: IOE produced the portion of the Virginia Graduation Competency Test which covers reading.

7. INTRAN

Louisiana: INTRAN assisted in item design for the portion of the Louisiana Assessment Program that deals with reading. The tests are currently administered in grades 4, 8, and 11. In 1982 this test will become a pass/fail test controlling movement to higher grades, starting with grade 2.

8. MCGRAW-HILL (CTB)

District of Columbia: CTB helped develop the customized Prescriptive Reading and Math Tests that are used in the District.

9. NATIONAL EVALUATION SYSTEMS

Connecticut: NES assisted in the development of the Connecticut Assessment of Educational Progress.

Georgia: NES is assisting in the development of a kindergarten test for spring 1980.

Hawaii: NES is assisting in the development of an item pool and item design for a competency test program. The test will be administered in the third grade in 1980-81 and will eventually be given in the sixth, eighth, and tenth grades.

Maryland: NES is assisting in the development of competency tests in mathematics, writing, citizenship, survival, and the world of work.

Massachusetts: NES assisted in the development of an item pool and selected a sample of communities for field testing of the Massachusetts Assessment of Basic Skills.

New Jersey: NES is assisting in the development of new items for the Minimum Basic Skills Test. This test is administered in grades 3, 6, 9, and 11 in reading and mathematics.

Rhode Island: NES assisted in item development for the Rhode Island Life Skills Test. This test is administered in the eleventh grade. The University of Rhode Island Curriculum and Research and Development Center holds the current contract.

Virginia: NES is assisting in item development and field testing for the Basic Learning Skills Test Program. The tests cover reading and mathematics and are given in grades 1 through 3. In 1980 the tests will be extended to grades 4 through 6.

10. NATIONAL TESTING SERVICE

Delaware: NTS is assisting in the development of an item pool that school districts may use in designing local competency tests. Local school districts must test but do not have to use the item pool.

Louisiana: NTS is assisting in item development for the writing and mathematics portions of the Louisiana Assessment Program. The tests are currently administered in grades 4, 8, and 11. In 1982 this will become a pass-fail test controlling movement to the higher grades, starting with grade 2.

11. NORTHWEST EVALUATION ASSOCIATION
(consortium of state and local school officials in Oregon and Washington).

Wisconsin: The Northwest Evaluation Association is assisting in the development of an item pool that school districts may use at their discretion.

12. NORTHWEST REGIONAL LABORATORY

Alaska: NRL assisted in item development for the Alaska Statewide Assessment.

Idaho: NRL assisted in item development for the Idaho Proficiency Test.

Oregon: NRL assisted in the development of an item pool and field testing for the Oregon Statewide Assessment.

13. RESEARCH MANAGEMENT CORPORATION
(part of UNCO in Washington, D.C.)

New Hampshire: RMC is currently assisting in item design for the New Hampshire Statewide Assessment Program.

14. RESEARCH TRIANGLE

Illinois: RT assisted in the development of the Illinois Inventory of Educational Progress. This test is used to provide sample assessments in reading, mathematics, and citizenship.

Maine: RT is assisting in the development of a test to replace the Maine Assessment of Educational Progress.

15. SCHOLASTIC TESTING SERVICES

North Carolina: STS produced two out of three of the currently used versions of the Minimal Competency Test.

Tennessee: STS assisted in the development of the Basic Skills Test. This competency test in reading, language arts, spelling, and mathematics is administered in the eighth grade.

Virginia: STS produced the mathematics component of the Virginia Graduation Competency Test.

16. SCIENCE RESEARCH ASSOCIATES (IBM)

Missouri: SRA developed the customized sixth-grade tests in reading and mathematics that are part of the Missouri Statewide Testing Program. All tests in this program are offered to school districts but are not required or used on a voluntary sample basis.

17. TOUCHSTONE APPLIED SCIENCE ASSOCIATES

Connecticut: TASA produced the Degrees of Reading Power, a test of reading proficiency or competence. In 1979-80 Connecticut used the test in the ninth grade to identify students who need remediation.

New York: TASA produced the Degrees of Reading Power and the Degrees of Writing Power. Both of these tests are part of New York's competency testing package. The Degrees of Reading Power attempts to determine what someone can read in the way of ordinary prose. It is currently given in the sixth, ninth, eleventh, and twelfth grades. In January 1981, passing this test will be a requirement for graduation. It will be administered three times a year to eleventh and twelfth graders, so that a student is given six chances to pass the test. The Degrees of Writing Power attempts to determine how well students can write, compared with predefined characteristics of good writing. The test is teacher-scored. It was administered in 1978 to ninth-grade students who were not in the Regents Program (college-bound track). It will eventually be administered on the same basis as the Degrees of Reading Power.

The Degrees of Reading Power was produced by TASA under contract with the New York State Board of Regents. Dr. Bertram Koslin, who once co-owned TASA, stated that the contract involved federal funds tapped by New York. However, the test is now jointly owned by the New York State Board of Regents and the College Examination Board, which is marketing the test. The College Board plans to buy out the Regents share and become the sole owner.

Appendix C

NEA'S ANALYSIS OF THE WIRTZ REPORT ON DECLINING SAT SCORES

EXECUTIVE SUMMARY

For more than one hundred years psychologists and educators have been using tests to measure human abilities. The 1880's were Galton's decade in the mental testing field, followed by Cattell (1890's) and Binet (1900's). Actually tests and measurement as they affect our life today evolved from at least three interrelated developments: (1) the study of individuals who deviated from the norm, (2) the experimental study of normal adult behavior, and (3) the development of mathematical models as tools for measurement.

More recently, the use of mental tests for sorting and selecting students by colleges and universities has become the work of the Educational Testing Service (ETS), which is a private, nonprofit organization devoted to measurement and research primarily in the field of education. It was founded in 1947 by the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board (CEEB).

Since 1972 ETS has had a budget of over $47 million, with a 1976 budget of $62.9 million. Testing activities amounted to $55.8 million of the revenue; the balance came from research, development, instructional services, and other. Actually, $2.9 million (4:6 percent) of ETS's revenue came from the federal government.

Objectives

The five-member NEA Special Committee on the (Wirtz) Report on Declining SAT Scores reviewed the charge from President Ryor and developed three objectives and nine related questions with which to analyze the CEEB-ETS report, as follows:

OBJECTIVE ONE: To analyze *On Further Examination,* the College Entrance Examination Board's report of the advisory panel on the Scholastic Aptitude Test score decline.

Question No. 1: What were the highlights of the CEEB-ETS report on the declining SAT scores?

Question No 2: What were the significant findings of the CEEB-ETS study about the SAT score decline?

Question No. 3: Is there any evidence in the CEEB-ETS report that the SAT should continue to be used by institutions of higher education as a standard to select students for admission?

Question No. 4: What were the implications of the CEEB-ETS report for classroom instruction?

OBJECTIVE TWO: To review NEA's current policy on standardized testing, considering the following: CEEB's *On Further Examination*, the attempt on the part of selected members of the U.S. Congress to pass federal legislation on testing, the related impact on local school district curricula, and teacher evaluation.

Question No. 5: What is NEA's current policy on standardized tests?

Question No. 6: Should NEA change its policy on standardized tests?

Question No. 7: What should NEA's position be on the "back to basics" controversy and on the attempts being made to reduce curricula offerings at the local level?

OBJECTIVE THREE: To develop a set of recommendations for presentation to the various levels of NEA governance and to alert standing committees of policy recommendations.

Question No. 8: In which areas are policy recommendations needed on testing?

Question No. 9: In which areas are recommendations needed to improve NEA program activities in the field of standardized testing?

CONCLUSIONS

Three objectives and nine questions were used by NEA's Special Committee on Declining SAT scores to analyze the College Entrance Examination Board's report *On Further Examination* and make recommendations. The objective and a brief statement of the Committee's conclusion about each objective are presented in this section.

- *OBJECTIVE ONE: To analyze* On Further Examination, *the College Entrance Examination Board's report of the advisory panel on the Scholastic Aptitude Test score decline.*

An analysis of the five sections and related studies included in the Wirtz report produced a mixed reaction about the findings. A substantial amount of evidence was presented in the form of descriptive statistics, which suggested that the study of the decline of the Scholastic Aptitude Test scores was not possible.

A comparison of just the number of students completing high school, entering college, and taking the SAT suggests that the last 25 years has produced not only more students to be educated but also a need for multiple criteria (standards), not just one criterion that applies to all students throughout the country. For the CEEB panel to have extended its study beyond the demographic data presented raises a question about what was expected to be found in all the isolated univariate type of studies that were commissioned and that appear in the appendixes to *On Further Examination.*

An analysis of the ETS auditors' report for 1975 and 1976 shows that revenue from testing activities was approximately $49 million in 1975 and $56 million in 1976. The SAT produced an estimated $9.1 million in 1975 and $9.8 million in 1976. In both years this equaled about 18 percent of ETS's annual revenue.

If only 7 percent of the 1976-77 high school graduates took the SAT—as was the case in 1951-52—there would be an $8.3 million reduction in ETS's revenue. Such a reduction of revenue would obviously have a significant impact on ETS's activities and staffing. The SAT is one of the corporation's greatest sources of income. For the Wirtz panel to have concluded anything about the SAT that would have produced less use of the test was highly unlikely.

ETS is not the only corporation earning a substantial amount of money from testing activities. Many book publishers profit from selling tests and books that help produce good test results. For example, Harcourt, Brace, and Jovanovich sells the Stanford Achievement Test and the OTIS Group Intelligence test; Houghton Mifflin markets the Stanford-Binet, Lorge-Thorndike, and Iowa Tests of Basic Skills; McGraw-Hill owns the California Test Bureau, which sells the California Achievement Test battery, and International Business Machines owns Science Research Associates.

In the Committee's opinion the problem is the unwillingness of the testing industry to apply contemporary technology to improve the state of the art in testing.

The Wirtz report more than adequately answers the questions about the reliability of the SAT and attempts to answer the question about predictive validity.

The more significant validity questions about construct validity (the underlying theoretical basis of what is actually being measured by the instrument, combined with supportive statistical and logical data from research studies) and content validity (which relates to the content currently being taught in the schools) were not adequately investigated or at least not reported.

To use the concept of an "unchanging standard" and to begin to investigate the changes in schools and society for 25, 20, or even five years do not suggest that the most objective approach was used to evaluate the decline in the SAT scores.

The Wirtz report provides a two-premise explanation about the 14-year SAT score decline.

The first premise portrays the decline for the first six or seven years as being caused by a markedly changing SAT-taking population. During this interval (1963 to 1969) there were "larger proportions of characteristically lower-scoring groups of students."

The second premise attributes the decline in the last seven years to "factors in the schools and in society at large." The changing nature of societal values caused the schools to attempt to provide a more diversified curriculum for the various groups of students who had not previously had the opportunity or need, in terms of employment, to reach high school or beyond.

The CEEB panel had to resort to explaining the score decline between 1970-77 to "circumstantial evidence." In Part Four of the report more than 50 theories were examined and discussed by the panel. Each of these theories held three assumptions in common: One, "that since the problem has been reduced to a single statistic—the drop in these averages—there must be a single answer; second, that what has happened is in every respect bad; and third, that whatever caused it is somebody else's fault."

The panel's "only certain conclusion is that we are dealing here with a virtually seamless web of causal connections. [The] most critical elements emerge more clearly in looking first at some developments in the schools, then at several major societal changes, and finally at the murky but probably vital area of youths' motivations."

Twenty-seven published appendixes were reported along with the findings of the CEEB SAT score decline. There was extensive use of descriptive statistics and studies with nonrandom sampling selection techniques. Nonrandom sampling restricts the panel's ability to make generalizations about students in all 50 states

and the 16,000-plus school districts. Instead, the panel was forced to make decisions based on isolated studies and what it termed "circumstantial evidence." Specifically, the conclusions relating to television were termed "essentially subjective."

The report had an overriding tone throughout about "traditional" standards and values, which were challenged (parenthetically) by limiting statements in the report. However, the statements of consensus provided only subjective, nongeneralized conclusions about the SAT score decline. In fact, the two types of score decline between the arbitrary 14-year interval of 1963-77 were attributed to "changing membership of the population tested" and "six other sets of developments."

The six other sets of developments were determined to have a beginning in 1971. It was acknowledged that the "forces" began before 1971; however, the effects can only be attributed to the 1971-77 interval.

Why there are six sets of developments rather than one, two, three, or nine is not adequately addressed in the report. Frequent reference is made to the dynamics of change in society and the historical consistency (reliability) of the SAT without any reference or question about the validity of the SAT as a surrogate for society and its unchanging standards.

It appears that the CEEB-ETS report could have been written by any panel charged with developing circumstantial evidence about the decline of the SAT scores over the past 40, 30, or 25 years. The studies that were used as a basis to reach conclusions do not provide a scientific data base on which to make an objective evaluation about the alleged decline in SAT scores.

NEA's Special Committee raised the question about the continued use of the SAT for selecting students for college admission. An analysis of continued use produced the following conclusion.

The SAT is considered to be a maximum performance test. It was designed to predict success in college. Tests of mastery of school subjects are called achievement tests. The SAT is an aptitude test and not an achievement test. The difference between achievement and aptitude tests is in the way in which they are used.

A test is generally referred to as an achievement test when it is used to determine a person's success in "past" study. The same test when used to forecast future success in a course or assignment is generally referred to as an aptitude test.

The way the test is used determines the classification of the test. Generally, when a test, such as the SAT, is used to predict future academic performance *prognosis*, it is classified as an aptitude test. When a test is used for diagnosis it is referred to as an achievement test.

Teachers use diagnostic (achievement) tests in their day-to-day teaching of students. The use of a test to analyze a student's performance on a set of tasks to improve learning is an appropriate use of tests.

The use of a test to reject or select an individual for college or employment tends to foster racism, elitism, classism, and separatism. The CEEB-ETS report provides many examples of how the SAT does discriminate against students who belong to the lower socioeconomic groups, minorities, and women. It is precisely for these types of reasons that the NEA is searching for a different means to measure achievement and to do away with aptitude tests.

The predictive validity of the SAT does not compare favorably to the grades given by a teacher as a predictor of future success in college. The CEEB-ETS reported findings of the studies on the predictive validity of a student's scores on the SAT and a student's high school grades were consistent with previous studies. Bloom and Peters reported a validity coefficient between high school and college grades of .5 dating as far back as 1926. The point to be made is that for more than 50 years, high school grades have been the best predictor of college grades. The use of the SAT adds very little to a college's ability to predict future success.

It would seem more rational to use both an achievement test *and* grades to determine a student's current ability to perform. At least this approach would help in the diagnosis and future development of the student.

In summary, it appears that the SAT cannot adequately predict a student's success in college. Furthermore, the changing needs of society, families, students, and teachers conflict with the "unchanging standard" that the makers of the SAT profess to have built into the test.

NEA's Special Committee reviewed other published articles and references about the SAT, including an article by Ralph W. Tyler, a member of the Wirtz panel, in which both he and Benajmin S. Bloom, another panel member, provide their own explanations about the score decline. Tyler commented about children's achievement and Bloom about the score decline. Their conclusions are as follows.

Tyler:

> The available data regarding the educational achievements of our children are not wholly consistent with the trend in Scholastic Aptitude Test scores. The National Assessment of Educational Progress, for example, furnishes information on the educational achievements of a reliable sample of nine-year-olds, thirteen-year-olds, seventeen-year-olds, and young adults, ages twenty-six to thirty-five. In a survey taken first in 1971, and again in 1975, National Assessment found that, nationwide, an estimated fifty thousand more nine-year-olds were able to respond correctly to a typical reading item in 1975 than in 1971. The reading performance of seventeen-year-olds has also improved somewhat during the past four years. On the other hand, reading achievements of thirteen-year-olds has changed little during this period.
>
> In mathematics, National Assessment found that ninety per cent of seventeen-year-olds can add, subtract, multiply, and divide accurately with whole numbers, but only forty-five per cent can use these computational skills properly in working out unit costs, the amount of income tax due, and other quantitative problems often encountered by adults. In science, in 1969-1970 when the mass media was emphasizing the importance of science, thirteen-year-olds and seventeen-year-olds performed five per cent better than four years later when science was given less favorable treatment in the press. In writing (composition), the average score of thirteen-year-olds and seventeen-year-olds has declined consistently.
>
> The Scholastic Aptitude Test data show that the decline has been greater in the verbal sections than in the mathematics ones, and has been greater in the sections testing vocabulary than in reading.

Bloom:

> I think there is a lot wrong with American education, but the Scholastic Aptitude Test is not where you are going to identify it. The S.A.T. comparative figures are based on the 1941 version of S.A.T., when approximately forty-one thousand students—most of them going to Ivy League colleges—took the test. Today, about two million young people are going to colleges, mostly public; about a million and a half take the S.A.T. test.
>
> The first major drop in S.A.T. scores took place between 1941 and 1951. By 1951, about half a million students were taking the tests, and many of them were heading for institutions other than Dartmouth or Swarthmore or Yale.
>
> From 1951 to 1977, the drop in the verbal score has been about fifty points and the drop in the mathematics score about thirty points. About twenty-eight of the fifty points in the verbal-score decline and twenty-three of the thirty points in the mathematics score are attributable to the change in the composition of the college population during that period. In 1951, more than half of the students who took the S.A.T. were in the upper twenty per cent of their high-school class. Today, about a third of the students taking the S.A.T. are in the upper twenty per cent of their class. The compositional change does not refer to blacks or Chicanos. It refers

for the most part to white children coming from different sectors of their high-school graduating class.

I should point out that the rest of the drop in the S.A.T. scores—that is, that which we cannot clearly account for—concerns three test items out of approximately ninety in the verbal test and two items out of approximately seventy-five or eighty in the mathematical test.

It is also important to note that the S.A.T. is a speed test. Almost no student can finish the test in the allotted time. If you were to let each student have as much time as he wanted, the distribution would be very different. At one time, students would prepare for several weeks—some of them would prepare for a year—getting ready for the S.A.T., developing speed in answering questions and solving problems. There is very little of that kind of preparation now. Also, students used to repeat the S.A.T. and increase their score by about thirty points. Today, students take the S.A.T., and whatever score they get, they let it stand. The number of students retaking the S.A.T. between their junior and senior year in high school has decreased enormously. In the minds of students, the importance of the S.A.T. as the major gatekeeper in American education has dropped significantly.

In addition, Tyler identifies a number of implications for the classroom and for society. There is a need for more writing assignments, the critical use of television as a supplementary resource in the learning process, and the examination of the out-of-school educational environment.

Finally, it is reassuring to know that there is no evidence to support the view that children are learning less today. There is a need to determine what and when society wants students to learn what is deemed valuable and important. If there is a need for writing assignments, there will have to be accommodations both within the school curricula and in the out-of-school experiences.

OBJECTIVE TWO: To review NEA's current policy on standardized testing, considering the following: CEEB's On Further Examination, *the attempt on the part of selected members of the U.S. Congress to pass federal legislation on testing, the related impact on local school district curricula, and teacher evaluation.*

The second objective was principally directed toward the review and examination of current NEA policy on standardized tests. After studying the report of NEA's Task Force on Testing, the Committee concluded that there was a need to rewrite the current resolution.

The proposed resolution appears in the recommendations.

OBJECTIVE THREE: To develop a set of recommendations for presentation to the various levels of NEA governance and to alert standing committees of policy recommendations.

The Committee developed three policy and five program recommendations. The recommendations appear on pages 53-55.

Appendix D

NEA 1980 RESOLUTIONS CONCERNING TESTING, CRITERION-REFERENCED TESTS, AND TRUTH IN TESTING

H-10. Testing

The National Education Association recognizes that testing of students, preschool through job entry, may be appropriate for such purposes as—

a. Identifying learning needs
b. Recommending instructional activities
c. Describing student progress.

The Association opposes the use of tests that deny students full access to equal educational opportunities, or that are used to evaluate teachers.

The Association believes that standardized tests should not be administered when they are—

a. Potentially damaging to a student's self-concept
b. Biased
c. Used as the only criterion for student placement
d. Invalid, unreliable, or out-of-date
e. Used as a basis for the allocation of federal, state, or local funds
f. Used by testing companies or publishers to promote their own financial interests at the expense of sound educational uses
g. Used to compare individual schools
h. Used in an exploitive manner by the media
i. Used as the sole criterion for graduation or promotion
j. Inappropriate for the use intended.

Revised resolution.

H-11. Criterion-Referenced Tests

The National Education Association believes that criterion-referenced tests are a viable alternative to standardized norm-referenced tests. Such tests should be designed to describe student performance based on carefully developed curriculum. It is inappropriate to administer criterion-referenced tests that do not specifically measure instructional content.

Staff, time, instructional materials, and other resources should be provided to assist students who experience difficulty achieving the desired criteria reflected by tests.

New resolution.

H-12. Truth-in-Testing

The National Education Association believes that intelligence, aptitude, and achievement tests have historically been used to differentiate rather than to measure performance and have, therefore, prevented equal educational opportunities for all students, particularly minorities, lower socioeconomic groups, and women. Contemporary research on the structure of the intellect identifies multiple and varied mental operations and advances the significant premise that these operations can be taught, that intelligence is dynamic rather than fixed.

The Association further believes that the truth-in-testing movement is an important step for bringing about long-needed test reform. Therefore, it urges all state affiliates to strive for passage of truth-in-testing legislation that includes a provision for each individual test taker to receive a copy of all test questions, scores, and rationale for correct answers.

New resolution.

Appendix E

NEA'S ANALYSIS OF H.R. 3564 AND H.R. 4949

Two legislative proposals concerning educational testing are before the Committee on Education and Labor. The first proposal, referred to as "Truth-in-Testing Act of 1979" (H.R. 3564), was introduced by Rep. Gibbons. The second proposal, the "Educational Testing Act of 1979" (H.R. 4949), was introduced by Rep. Weiss. The latter proposal (H.R. 4949) is based on New York legislation proposed and passed during the summer of 1979.

H.R. 3564 and H.R. 4949 concern the use of standardized tests, a subject about which NEA has raised questions and expressed concerns. Because of the NEA concern with the use of standardized tests, both proposals have been analyzed in terms of their similarities, their differences, and their responsiveness to NEA concerns.

In general, NEA believes that the two proposals represent somewhat different approaches to the use of standardized tests. To the extent that H.R. 3564 and H.R. 4949 are responsive to NEA concerns, both proposals should be supported. The Gibbons "Truth-in-Testing Act" (H.R. 3564), however, is expected to generate more opposition in Congress and could, if passed, prove to be a less successful vehicle for meeting the concerns expressed by NEA.

Both H.R. 3564 and H.R. 4949 represent notice and disclosure legislation. They differ substantially as to the type of tests covered, the extent of involvement of the Commissioner of Education and the type of enforcement provisions. H.R. 3564 would cover the National Teacher Examination which is a concern of NEA. The bill would also cover other occupational tests, which will engender opposition, and tests other than standardized tests, regulation of which would probably prove unworkable. For the most part, the disclosure requirements of H.R. 3564 require the type of information currently provided voluntarily by testing agencies such as ETS. Because H.R. 3564 does not require disclosure of underlying data on the examinations, it would not enable professionals outside the testing industry, including teachers, to analyze or comment on test construction and validity. In addition, H.R. 3564 fails to provide for disclosure of scoring data in addition to test scores which may be given to educational institutions. Groups favoring testing disclosure laws have stated that testing agencies provide information such as suspicions of cheating, unacknowledged repetition of a test and factors based on current school attended to be used in evaluating the score. Students have not been informed of this type of information where it is incorrect. In addition, students have not been provided with their test answers and the correct answers. Thus, students have been unable to learn of or correct computer grading errors. The Gibbons bill does not address this problem either.

In contrast, each one of the instances noted above is addressed in the Weiss bill with the exception of occupational testing. Various portions of the Weiss bill could use clearer and better language. In addition, some consideration should be given to the viability of including financial regulation of the testing companies in this legislation.

In addition to standardized tests, H.R. 3564 covers "oral" tests, "practical" tests and "demonstration" examinations. Sec. 2(3). The bill apparently reaches practical or demonstration examinations used in occupational licensing such as barbering, oral examinations such as the foreign service examinations, and practical or demonstration examinations used in educational admissions such as submission of a portfolio to an art school or a stage performance required for a drama school. Regulation of such tests would probably be unworkable.

H.R. 3564 contains both pre-test (Sec. 6(a)) and post-test (Sec. 6(b)) disclosure requirements which require information to be provided to test takers. Prior to administration of the test, each applicant must be provided with a written notice containing essentially the types of information currently provided voluntarily by the testing companies:

1. A detailed description of the area of knowledge or the type of aptitude that the test attempts to analyze;
2. In the case of a test of knowledge, a detailed description of the subjects to be tested;
3. The margin of error or the extent of reliability of the test, determined on the basis of experimental uses of the test and, where available, actual usage;
4. The manner in which the test results will be distributed by the testing entity to the applicant and to other persons; and
5. A statement of the applicant's [post-test notification] rights.

The post-test notification provision requires that "promptly upon completion of scoring" the test taker must be notified of:

1. The individual's specific performance in each of the subject or aptitude areas tested;
2. How that specific performance ranked in relation to the other individuals and how the individual ranked on total test performance;
3. The score required to pass the test for admission to such occupation or the score which is generally required for admission to institutions of higher education;
4. Any further information which may be obtained by the individual on request.

Section 6(c), the final substantive provision of the bill, prohibits the scoring of achievement tests on the basis of a curve:

> c. No educational or occupational admissions test which tests knowledge or achievement (rather than aptitude) shall be graded (for purposes of determining the score required to pass the test for admission) on the basis of the relative distribution of scores of other test subjects.

The enforcement provisions of H.R. 3564 (Sec. 7) authorize private causes of action by an aggrieved individual "whenever any person has administered or there are reasonable grounds to believe that any person is about to administer any test in violation of this act." The bill specifically provides for "preventive relief" including a permanent or temporary injunction and restraining orders and for appointment of counsel "in such circumstances as the court may deem just." The bill authorizes attorney's fees (Sec. 7(b)) and provides for federal court proceedings without regard to exhaustion of remedies. Sec. 7(c).

The enforcement procedures of injunction or restraining order represent onerous remedies, and it seems doubtful that federal courts will be inclined to enjoin the administration of standardized tests such as the SAT. For this reason, the remedies provided by the bill appear to be ineffective. Since the bill specifically authorizes "a civil action for preventive relief" courts may find that such relief is the exclusive remedy for violations of the Act.

The "Educational Testing Act of 1979" (H.R. 4949) identifies three legislative purposes (Sec. 2(b)):

1. To ensure that test subjects and persons who use test results are fully aware of the characteristics, uses, and limitations of standardized tests in postsecondary education admissions;
2. To make available to the public appropriate information regarding the procedures, development, and administration of standardized tests; and
3. To protect the public interest by promoting more dependable knowledge about the limits of appropriate usage of standardized test results and by promoting greater accuracy, validity, and reliability in the development, administration, and interpretation of standardized tests.

This bill requires more extensive pre-test disclosure to test takers than H.R. 3564 and, unlike H.R. 3564, specifically requires that the pre-test notice be provided contemporaneously with the test registration form. Sec. 3(a). The legislation specifically addresses the coachability issue and requires testing agencies to inform individuals of the extent to which their scores may be improved by taking a preparation course. Pre-test notice must include the following information:

1. The purposes for which the test is constructed and is intended to be used.
2. The subject matters included on such test and the knowledge and skills which the test purports to measure.
3. Statements designed to provide information for interpreting the test results, including explanations of the test, and the correlation between test scores and future success in schools and, in the case of tests used for post baccalaureate admissions, the correlation between test scores and success in the career for which admission is sought.
4. Statements concerning the effects on and uses of test scores, including—
   a. if the test score is used by itself or with other information to predict future grade point average, the extent, expressed as a percentage, to which the use of this test score improves the accuracy of predicting future grade point average, over and above all other information used; and
   b. a comparison of the average score and percentiles of test subjects by major income groups; and
   c. the extent, if available to the test agency, to which test preparation courses improve test subjects' scores on average, expressed as a percentage.
5. A description of the form in which test scores will be reported, whether the raw test scores will be altered in any way before being reported to the test subject, and the manner, if any, in which the test agency will use the test score (in raw or transformed form) by itself or together with any other information about the test subject to predict in any way the subject's future academic performance for any postsecondary educational institution.

6. A complete description of any promises or covenants that the test agency makes to the test subject with regard to accuracy of scoring, timely forwarding or score reporting, and privacy of information (including test scores and other information)[1], relating to the test subjects.
7. The property rights of the test subject to the test results, if any, the duration for which such results will be retained by the test agency, and policies regarding storage, disposal, and future use of test scores.
8. The date by which the test subject's test scores will be completed and mailed to the test subject.
9. A description of special services to accommodate physically handicapped test subjects.

In addition to providing notice to test subjects, the bill requires the testing agency to provide the same information to the recipient institution prior to or coincident with the reporting of test scores.

The major area covered by the Weiss bill is reporting to governmental educational agencies. Two types of information must be disclosed to the government. First, this reporting requirement concerns the studies and evaluations of the tests themselves and is designed to allow professionals outside the testing industry, including teachers, access to such studies to allow independent analysis of the construction, validity and use of the tests. The second type of information to be disclosed includes the test questions and answers and scoring rules. This is accomplished by cross-reference to the Freedom of Information Act, 5 U.S.C. Sec. 552(a)(3), which authorizes release of records. The test agency is required to provide to the Commissioner of Education:

- Any study, evaluation, or statistical report pertaining to a test, which a test agency prepares or causes to be prepared, or for which it provides data. Nothing in this paragraph shall require submission of any reports or documents containing information identifiable with any individual test subject. Such information shall be deleted or obliterated prior to submission to the Commissioner, [and]
    1. shall, within 30 days after the results of any standardized test are released, file or cause to be filed in the office of the Commissioner—
        a. a copy of all test questions used in calculating the test subject's raw score;
        b. the corresponding acceptable answers to those questions; and
        c. all rules for transferring raw scores into those scores reported to the test subject and postsecondary educational institutions together with an explanation of such rules.

This data, in addition to being made available pursuant to the Freedom of Information Act, must be made available by the Commissioner of Education to state educational agencies and commissions.

The testing agency must also provide the questions, the correct answers, and the test taker's answers, as well as scoring information, to the test subject on request for a 90-day period subsequent to release of the test scores.

Furthermore, the legislation requires the Commissioner of Education to prepare for Congress an evaluation of the data on these tests both with regard to coachability and cultural bias:

b. The Commissioners [sic] shall report to Congress within one year of the effective date of this Act concerning the relationship between the test scores of test subjects and income, race, sex, ethnic, and handicapped status. Such report shall include an evaluation of available data concerning the relationship between test scores and the completion of test preparation courses.

The major difference between the Weiss draft and the New York law upon which it is based is an attempt in the federal legislation to regulate the costs to test subjects of the tests and to require financial disclosures by the testing companies. During the New York hearings, the testing companies argued that costs would skyrocket. Proponents of the New York legislation, New York Public Interest Research Group and Nader in particular, questioned these predictions using whatever data they could obtain from the testing companies, especially ETS. Section 7 of the bill entitled "Testing Costs and Fees to Students" provides as follows:

In order to ensure that tests are being offered at a reasonable cost to test subjects, each test agency shall report the following information to the Commissioner:

1. Before March 31, 1981, or within 90 days after it first becomes a test agency, whichever is later, the test agency shall report the closing date of its testing year. Each test agency shall report any change in the closing date of its testing year within 90 days after the change is made.
2. For each test program, within 120 days after the close of the testing year the test agency shall report:
   a. the total number of times the test was taken during the testing year;
   b. the number of test subjects who have taken the test once, who have taken it twice, and who have taken it more than twice during the testing year;
   c. the number of refunds given to individuals who have registered for, but did not take, the test;
   d. the number of test subjects for whom the test fee was waived or reduced;
   e. the total amount of fees received from the test subjects by the test agency for each test program for that test year;
   f. the total amount of revenue received from each test program, and
   g. the expenses to the test agency of the tests, including:
      (1) expenses incurred by the test agency for each test program;
      (2) expenses incurred for test development by the test agency for each test program; and
      (3) all expenses which are fixed or can be regarded as overhead expenses and not associated with any test program or with test development;
3. If a separate fee is charged test subjects for admissions data assembly services or score reporting services, within 120 days after the close of the testing year, the test agency shall report:
   a. the number of individuals registering for each admissions data assembly service during the testing year;
   b. the number of individuals registering for each score reporting service during the testing year;

c. the total amount of revenue received from the individuals by the test agency for each admissions data assembly service or score reporting service during the testing year; and

d. the expenses to the test agency for each admissions data assembly service or score reporting service during the testing year.

The Weiss bill, like the New York legislation, uses a civil penalty as its remedy. While the New York law establishes a $500 penalty per violation, the federal law establishes a $2,000 fine. This would represent a small penalty where the test agency failed to properly report to the Commssion of Education since this would probably constitute a single violation. With regard to violations of the notice to students the penalties could be substantial since presumably failure to provide the required notices to students would result in multiple violations reflecting the number of students involved. One potential difficulty in enforcement may be in determining which and how many individuals were not given proper notice or timely reporting. The Commissioner is authorized by the draft to promulgate regulations to implement the legislation and enforcement would be one area where regulations might fill in the sketch created by the draft legislation.

The Weiss bill would require disclosure to students of covenants and promises made by the testing agencies. Private causes of action by test takers could be based on breaches of these contractual warranties.

FOOTNOTE

[1]The phrase "and other information" was added by Weiss's staff subsequent to conversations with NEA. Significant questions exist as to the use made by ETS of personal data obtained on the test or test application. ETs sells student lists to institutions.

Appendix F

nea

EXECUTIVE OFFICE

NATIONAL EDUCATION ASSOCIATION • 1201 16th St., N.W., Washington, D C 20036 • (202) 833-4000

JOHN RYOR, President
WILLARD H. MCGUIRE, Vice-President
JOHN T. MCGARIGAL, Secretary-Treasurer

TERRY HERNDON, Executive Director

June 8, 1979

Dr. Warren G. Hill
Executive Director
Education Commission of the States
1860 Lincoln Street
Denver, Colorado 80203

Dear Dr. Hill:

The National Education Association strongly supports the Education Commission of the States' application to continue as the organization responsible for the National Assessment of Educational Progress. The National Assessment has gained respect from teachers, administrators, and educational policy makers at all levels of the education community over the last fifteen years.

NEA advocates measurement techniques and approaches which help policy makers formulate intelligent decisions about school programs. The National Assessment has provided this information in the past and it is hoped that the program can be extended down into the local school districts to replace the current fad of competency testing.

NEA strongly supports the makeup of the National Assessment Policy Committee which includes teacher representation on the committee. The Association would urge that the Federal Government continue this practice and require that teachers be represented on the National Assessment Policy Committee in direct proportion to their national membership. The four teachers on the committee should be designated by the majority organization or, if this is not possible, allocated to teacher organizations according to membership. Administrator and school board organizations should designate their representatives.

The NEA recommends that the ECS be granted the funds to continue the NAEP.

Sincerely,

Terry Herndon
Executive Director